(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 July 2001 (19.07.2001)

PCT

(10) International Publication Number
WO 01/51664 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/US01/01081

(22) International Filing Date: 12 January 2001 (12.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/175,669 12 January 2000 (12.01.2000) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/175,669 (CIP)
Filed on 12 January 2000 (12.01.2000)

(71) Applicant *(for all designated States except US)*: **DANA-FARBER CANCER INSTITUTE, INC.** [US/US]; 44 Binney Street, Boston, MA 02115 (US).

(71) Applicant *(for US only)*: **PARDEE, Arthur** (heir of the deceased inventor) [US/US]; 30 Codman Road, Brookline, MA 02445 (US).

(72) Inventor: **SAGER, Ruth** (deceased).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **MARTIN, Katherine, J.** [US/US]; 61 Carleton Road, Belmont, MA 02478 (US). **PARDEE, Arthur** [US/US]; 30 Codman Road, Brookline, MA 02445 (US).

(74) Agent: **ELRIFI, Ivor, R.**; Mintz Levin Cohn Ferris Glovsky and Popeo P.C., One Financial Center, Boston, MA 02111 (US).

(81) Designated States *(national)*: CA, JP, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/51664 A2

(54) Title: METHOD OF DETECTING AND CHARACTERIZING A NEOPLASM

(57) Abstract: Disclosed are methods of detecting neoplasms, such as breast carcinomas, using differentially expressed genes. Also disclosed are methods of identifying agents for treating neoplasms.

METHOD OF DETECTING AND CHARACTERIZING A NEOPLASM

FIELD OF THE INVENTION

The invention relates to methods of detecting cancer.

BACKGROUND OF THE INVENTION

Cancer is a morphologically and genetically heterogeneous disease. For a given cancer, no simple relationship may exist between a mutation in a gene, the expression level of the gene, and a certain etiology or extent of the disease. Consequently, prognostic marker systems based on single parameters can be inadequate to assess the disease.

A detailed molecular characterization or fingerprint of cancer is an objective recently made possible by the development of several new high throughput analytical methods. These include techniques for the analysis of DNA, mRNA, and proteins within a cell. A goal that is now in sight is to build databases of detailed molecular information and to link them to clinical information. This approach has the potential to help patients by accurately grouping tumor subtypes. Such categorization may enable clinicians to more accurately distinguish prognostic groups and to predict the most effective therapies. For example, new prognostic methods may more accurately predict the benefit to patients of chemotherapy treatments.

SUMMARY OF THE INVENTION

The invention is based the discovery of that the pattern of expression of clusters of genes are correlated to a cancerous state. Accordingly, the invention features a method of diagnosing a neoplasm in a subject by (a) measuring a subject profile of tumor-associated genes; and (b) comparing said subject profile with a reference profile of said tumor-associated genes. A difference between the subject profile of expression of a tumor-associated genes and the subject profile of expression of tumor associated genes indicates that the subject suffers from or is at risk of developing a neoplasm. A profile of gene expression is a pattern of the level of expression of at least two tumor-associated genes. By tumor associated gene is meant a gene the level of expression of which differs in a cancer cell compared to a normal (or non-cancer) cell.

A reference profile is a single expression pattern derived from a single reference population or from a plurality of expression patterns. For example, the reference cell population

can be a database of expression patterns from previously tested cells for which one of the herein-described parameters or conditions (*e.g.*, estrogen receptor status or tumor stage) is known.

A non-cancer reference profile is determined by measuring the level of expression of tumor-associated genes in a non-cancer reference cell population. The non-cancer reference cell population is made up of substantially of noncancerous cells. For example, the population is at least 85%, preferably 95% and more preferably 99% and most preferably 100% non-cancerous cells as determined by a method which is not based on expression of DFCIs: 1-52. Similarly, a cancer reference profile is determined by measuring the level of expression of tumor-associated genes in a cancer reference cell population. A cancer reference cell population made up substantially cancerous cells. For example, the population is at least 85%, preferably 95% and more preferably 99% and most preferably 100% cancerous cells as determined by a method which is not based on expression of DFCIs: 1-52.

In another aspect, the invention includes a methods of diagnosing a neoplasm in a subject by (a) measuring expression of two or more DFCIX nucleic acid sequences in a subject derived cell population; to yield a subject profile; and (b) comparing the expression of said nucleic acid sequences to the expression of nucleic acid sequences in a cancer reference profile. A substantial similarity between the expression of nucleic acid sequences in said subject-derived cell population and the cancer reference profile indicates the presence of a neoplasm in the subject.

The invention also features a method of assessing the prognosis of a subject with a neoplasm, by (a) measuring over time the expression two or more DFCIX nucleic acid sequences in a subject derived cell population to yield a subject profile to yield a subject profile; and(b) comparing said subject profile to a cancer reference profile. An increase in similarity between said subject profile and said cancer profile over time indicates an adverse prognosis the subject.

In another aspect, the invention provides a method of assessing estrogen receptor status in a subject. The expression of one or more of the DFCIs:1-28 nucleic acid sequences in the test cell population is measured and compared to the expression of the nucleic acid sequences in a reference cell population, that includes at least one cell whose estrogen receptor status is known. By comparing the expression patterns in the test and the reference nucleic acid population, the estrogen receptor status in the subject can be determined.

Also provided in the invention is a method of assessing breast tumor stage in a subject. The method includes providing a test cell population from the subject. The expression of one or more of the DFCIs: 1-31 nucleic acid sequences in the test cell population is measured and

compared to the expression of the nucleic acid sequences in a reference cell population comprising at least one cell whose breast tumor stage is known. By comparing the expression patterns in the test and the reference nucleic acid population, the stage of a tumor in a subject can be determined.

Also within the invention is a method of assessing breast tumor size in a subject. The method includes providing a test cell population from the subject. The expression of one or more of the DFCIs: 29-35 nucleic acid sequences in the cell population is measured and compared to the expression of the nucleic acid sequences in a reference cell population, that includes at least one cell whose breast tumor size is known. By comparing the expression patterns in the test and the reference nucleic acid population, the size of a tumor in a subject can be determined.

Also included in the invention is a method of assessing the efficacy of a treatment of a neoplasm in a subject by (a) measuring the expression two or more of DFCIs: 1-51 and 52 nucleic acid sequences in a subject derived cell population to yield a subject profile; and (b) comparing the subject profile to a cancer reference profile. An increase in similarity between said subject profile and said cancer profile over time indicates the treatment is not efficacious. Conversely, a decrease in similarity between said subject profile and said cancer profile over time indicates the treatment is efficacious.

In another aspect, the invention includes a method for identifying a therapeutic agent suitable for treating a neoplasm in a selected subject. The method includes providing a test cell population from the subject. After the test cell population is contacted with a therapeutic agent, the expression of one or more of the DFCIX nucleic acid sequences in the test cell population is measured and compared to the expression of the nucleic acid sequences in a reference cell population that includes at least one cell whose neoplastic state is known. By comparing the expression patterns in the test and the reference nucleic acid population, a therapeutic agent suitable for the treatment of a neoplasm in a subject can be identified.

In a further aspect, the invention includes a method of identifying a candidate therapeutic agent suitable for treating a neoplasm. The test cell population is exposed to a test agent, and expression of the DFCIX nucleic acid sequences in the test cell population is measured and compared to the expression of the nucleic acid sequences in a reference cell population that includes at least one cell whose neoplastic state is known. By comparing the expression patterns in the test and the reference nucleic acid population, a candidate therapeutic agent suitable for treating a neoplasm in a subject can be identified.

In another aspect, the invention provides a method of categorizing a neoplasm in a subject. The expression of one or more of the DFCIX nucleic acid sequences in the test cell population is measured and compared to the expression of the nucleic acid sequences in a reference cell population that includes at least one cell whose neoplastic state is known. By comparing the expression patterns in the test and the reference nucleic acid population, the category of the neoplasm in a subject can be determined.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description, and from the claims.

DETAILED DESCRIPTION

The expression profile of five clusters of genes were found to be associated with four parameters used to characterize breast tumors. These parameters include estrogen receptor (ER) status, tumor stage, tumor size and tumor dissemination. By measuring expression of members of the various clusters of genes in a sample of cells, one or more of these parameters can be determined in a cell or population of cells. Similarly, by measuring the expression of members of these clusters in response to various agents, agents for treating cancer can be identified.

The genes whose expression levels indicated parameters associated with breast tumors are summarized in Table 1 and are collectively referred to herein as "DFCIX nucleic acids" or "DFCIX polynucleotides" and the corresponding encoded polypeptides are referred to as "DFCIX polypeptides" or "DFCIX proteins." Unless indicated otherwise, "DFCIX" is meant to refer to any of the sequences disclosed herein. (*e.g.*,DFCI 1-52),. For some genes, an increase in expression is indicative of a tumor parameter. For other genes, a decrease in expression is indicative of a tumor parameter. Some of the genes have been previously described and are presented along with a database accession number. Other genes are newly described (*e.g,* DFCI

7, 8, 25 and 31). For some genes, an increase or decrease in expression is indictive of more than one parameter. For example, an increase or decrease of maspin expression (DFCI 28) is indicative of ER status and tumor dissemination.

Table I. Genes Differentially Expressed In Cancerous and Non-Cancerous Breast Cells

DFCIX Sequence #	SEQ ID NO	Name	Genbank Accession No.
ER status-associated cluster I			
1		Keratin 19	Y00503
2		HE1	A18921
3		Unknown 31g	N27053
4		Unknown T31	AF063605
5		Proliferation Associated Gene	NM002574
6		108	AI860035
7	1	111f	None
8	2	109b	None
9		101f	AI866917
10		Protein Down Regulated in Cancer	AF081497
11		Condroitin sulf prot	NM004385
12		HER2/Neu Receptor	S57296
13		Ribosomal protein P2	M17887
14		Aldose Reductase-Like Protein	NM004812
15		p53	NM000546
16		unknown 36d	R79826
17		unknown 94	N39143
18		CD59	M84349
19		N-ras	X02751
ER status-associated cluster II			
20		Alpha 1-antichymotrypsin	J05176
21		Unknown JZ-mys	R96369
22		desmoglein 2	Z26317
23		histone H4	M16707
24		elafin	L10343
25	3	unknown T10	None
26		beta tubulin	J00314
27		CC3	K02765
28		maspin	U04313
Clinical stage-associated cluster			
29		unknown 47h	T91143
30		HSP-90	X15183
31	4	unknown 82(e)	None
Tumor size-associated cluster			
32		keratin 14	J00124
33		CD44	X62739
34		keratin 5	NM000424

35		GST pi	X06547
Disseminated tumor associated cluster I			
36		SRP19	Xm003959
33		CD44	X62739
37		TSP-2-8b	AJ132932
28		Maspin	U04313
38		HSIX1	X91868
39		Gro alpha	J03561
40		Myosin Light chain	U26162
41		Mdm-2	NM002392
42		ZZ38	None
26		Beta tubulin	AJ292757
43		N33 gene	U42360
44		Laminin alpha 3	L34155
Disseminated tumor independent genes			
45		Her2/neu	X03363
46		Beta actin	XM004814
47		Doc-1	NM004642
48		Mac25	L19182
49		Unknown 28/13	NM018579
50		Unknown TG90D	M22382
22		Desmoglein	Z26317
51		c-fos	K0650
52		Interferon γ	V00543
11		Chondrotin sulfate proteoglycan	J02814

None = Sequence does not match any sequence reported in Genbank

The invention involves measuring the expression of at least two, and up to all the DFCIX sequences listed in Table I. Using sequence information provided by the GeneBank database entries for the known sequences, or the sequence information for the newly described sequences, expression of the DFCIX sequences can be detected and measured using techniques well known to one of ordinary skill in the art. For example, sequences within the sequence database entries corresponding to DFCIX sequences, or within the sequences disclosed herein, can be used to construct probes for detecting DFCIX RNA sequences in, *e.g.*, northern blot hybridization analyses. As another example, the sequences can be used to construct primers for specifically amplifying the DFCIX sequences in, *e.g*, amplification-based detection methods such as reverse-transcription based polymerase chain reaction.

Expression level of one or more of the DFCIX sequences in the test cell population is then compared to expression levels of the some sequences in one or more cells from a reference cell population. The reference cell population includes one or more cells for which the compared

parameter is known, *e.g.*, estrogen receptor status, tumor size, tumor stage, neoplastic state (*i.e.*, the cell is cancerous or noncancerous) or whether the tumor has disseminated from the primary tumor site (*e.g.*, metastatic state). Whether or not the gene expression levels in the test cell population compared to the reference cell population reveals the presence of the measured parameter depends upon on the composition of the reference cell population. For example, if the reference cell population is composed of non-cancerous cells, a similar gene expression level in the test cell population and reference cell population indicates the test cell population is noncancerous. Conversely, if the reference cell population is made up of cancerous cells, a similar gene expression profile between the test cell population and the reference cell population that the test cell population includes cancerous cells.

A DFCIX sequence in a test cell population matches or is substantially similar in expression level to the expression level of the reference DFCIX sequence if its expression level varies within a factor of 2.0, 1.5, or 1.0 fold to the level of the DFCIX transcript in the reference cell population. In various embodiments, a DFCIX sequence in a test cell population can be considered altered in levels of expression if its expression level varies from the reference cell population by more than 1.0, 1.5, 2.0 or more fold from the expression level of the corresponding DFCIX sequence in the reference cell population.

If desired, comparison of differentially expressed sequences between a test cell population and a reference cell population can be done with respect to a control nucleic acid whose expression is independent of the parameter or condition being measured. For example, a control nucleic acid is one which is known not to differ depending on the cancerous or non-cancerous state of the cell. Expression levels of the control nucleic acid in the test and reference nucleic acid can be used to normalize signal levels in the compared populations. Control genes can be, *e.g.*, β-actin, glyceraldehyde 3- phosphate dehydrogenase or ribosomal protein P1 (36B4).

In some embodiments, the test cell population is compared to multiple reference cell populations. Each of the multiple reference populations may differ in the known parameter. Thus, a test cell population may be compared to a second reference cell population known to contain, *e.g.*, tumorous cells, as well as a second reference population known to contain, *e.g.*, non-tumorous cells.

In some embodiments, the test cell will be included in a cell sample from a subject known to contain, or to be suspected of containing, tumorous cells. In other embodiments, the cell sample will be derived from a subject from a region known to contain, or suspected of containing, a metastasis of a primary tumor, such as a breast carcinoma.

The test cell is obtained from a bodily fluid, *e.g.,* biological fluid (such as blood, serum, urine, saliva, milk, ductal fluid, or tears). For example, the test cell is purified from blood or another tissue. For many applications, *e.g.*, in assessing estrogen receptor status, assessing the efficacy of treatment, or in diagnosing a neoplasm in a subject, cells present in a bodily fluid can be examined instead of a primary lesion. Thus, the need for taking a biopsy from a known or suspected primary tumor site is obviated.

Preferably, cells in the reference cell population are derived from a tissue type as similar to test cell, *e.g.*, breast tissue. In some embodiments, the reference cell is derived from the same subject as the test cell, *e.g.*, from a region proximal to the region of origin of the test cell. In other embodiments, the control cell population is derived from a database of molecular information derived from cells for which the assayed parameter or condition is known.

The subject is preferably a mammal. The mammal can be, *e.g.*, a human, non-human primate, mouse, rat, dog, cat, horse, or cow.

The expression of 2, 3, 4, 5, 6, 7,8, 9, 10, 15, 20, 25, 28, 30, or 35 or more of the sequences represented by DFCI: 1-52 is measured and if desired, expression of these sequences can be measured along with other sequences whose level of expression is known to be altered according to one of the herein described parameters or conditions, *e.g.*, estrogen receptor status.

Expression of the genes disclosed herein can be measured at the RNA level using any method known in the art. For example, northern hybridization analysis using probes which specifically recognize one or more of these sequences can be used to determine gene expression. Alternatively, expression can be measured using reverse-transcription-based PCR assays, *e.g.,* using primers specific for the differentially expressed sequences.

Expression can be also measured at the protein level, *i.e.*, by measuring the levels of polypeptides encoded by the gene products described herein. Such methods are well known in the art and include, *e.g.*, immunoassays based on antibodies to proteins encoded by the genes.

When alterations in gene expression are associated with gene amplification or deletion, sequence comparisons in test and reference populations can be made by comparing relative amounts of the examined DNA sequences in the test and reference cell populations.

Assessing estrogen receptor status

The level of expression of some of the DFCIX sequences described herein is correlated with estrogen receptor (ER) status. Estrogen receptor status refers to the number of estrogen receptors present in a cancerous cell, *e.g.*, a breast cancer cell. Significant levels of estrogen receptors are associated with a favorable prognosis for the cancer (*See, e.g.*, Esteben et al., Cancer 74:1575-83, 1994).

The method includes providing a cell from the subject and detecting the expression level of one or more of the nucleic acid sequences DFCI: 1-28 in the cell.

Expression of the sequences is compared to a reference cell population. In general, any reference cell population can be used, as long as the ER status of the cells in the reference cell population is known. The reference cell population is made up substantially, or preferably exclusively, of noncancerous cells. An example of a noncancerous cell is the breast epithelial cell line MCF10A, which is ER-negative. Expression of DFCI: 1-19 is high in ER-positive cells and low in ER-negative tissues. Thus, high levels of DFCI: 1-19 in a test cell population compared to a reference cell population of a non-cancerous tissue such as the MCF10A cell line indicates the test cell population has high ER status.

Conversely, expression of sequences DFCI: 20-28 is higher in ER-negative tissues such as the cell line MCF10A. Accordingly, low levels of expression of sequences of DFCI: 20-28 in the test cell sample and compared to a reference cell sample including an ER-negative cell such as the MCF10A cell line indicates the sample has a high ER status. High levels of DFCI: 20-28 cells indicate the test sample has low ER status.

Alternatively, the reference cell population is made up substantially, or preferably exclusively, of cancerous cells. An example of a cancerous cell line is MDA-MB435, which has low ER status, and which has low levels of expression of DFCI: 1-19 and high levels of expression of DFCI: 20-28. For reference cell populations containing primarily or exclusively cells with low ER status, expression of DFCI: 1-19 is low and expression of DFCI: 20-28 is high. Thus, comparably low levels of expression of DFCI: 1-19 sequences, or comparably high levels

of expression of DFCI: 20-28 in a test cell population and the reference cell population indicates the test cell population has cells with low ER status. High levels of DFCI: 1-19 sequences or low levels of the DFCI: 20-28 sequences in the test cell population relative to cells in the latter reference population indicates the cells in the test population have high ER status. If desired, relative expression levels within the test and reference cell populations can be normalized by reference to expression level of a nucleic acid that does not vary according to ER status.

Assessing tumor stage

In another aspect, the invention provides a method of assessing tumor stage, *e.g.*, breast tumor stage, in a subject by comparing levels of DFCIX sequences in a test and reference population. Generally, tumor stage defines a period in the course of the disease. Breast tumors are staged by taking into account information on tumor size, nodal status, and distant metastases. Fisher, et al., Cancer Medicine, 3rd Ed., Eds: Holland et al., Vol. 2, ch. 32, pp. 1706-1750 (Lea & Febiger, Philadelphia), 1993. A more advanced stage indicates that the disease has progressed.

To assess tumor stage, a test cell population is taken from the subject previously diagnosed with a tumor and at least one of the DFCIX sequences are measured. DFCI: 29-31 are significantly over-expressed in stage IV breast tumors relative to stage I, II and III tumors. Accordingly, increased expression of one or more of these sequences in a test cell relative to a control cell indicates the tumor cell is from an advanced stage tumor. In other embodiments, altered expression of one or more of the sequences DFCIs: 1-28 indicates the stage of the tumor.

Levels of the DFCIX sequences in the test cell population are compared to the expression of the nucleic acid sequences in a reference cell population. The reference cell population includes cells whose tumor stage is known, or includes cells that are not cancerous. A similarity in expression patterns between the DFCIX sequences in the test sample and the pattern of expression of the reference sample indicates the test cell sample contains cells from a tumor that is of the same stage as those in the reference sample. Conversely, a difference in expression patterns between the test cell sample and control cell sample indicates the test cell sample is not from the same stage as the tumor cell or cells in the reference cell sample.

If desired, expression of these sequences can be measured along with expression level of other sequences whose expression is known to be altered according to tumor stage.

Assessing tumor size

The invention also includes a method of assessing tumor size, *e.g.*, breast tumor size, in a subject. Tumor size is an important independent predictor of disease prognosis. Fisher, B. et al., pp. 1706-1750 (Lea & Febiger, Philadelphia), 1993. Generally, a larger tumor size indicates a less favorable prognosis.

Tumor size is assessed by measuring expression of one or more DFCIX nucleic acid sequences in a test cell population within, or taken from, a subject. The test cell population includes at least cell known to be, or suspected of being, cancerous. The expression of the DFCIX nucleic acid sequences in the test cell population is compared to the expression of the same DFCIX nucleic acid sequences in a reference cell population. The reference cell population includes at least one cell whose breast tumor size is known, or which does not include a tumor cell.

Similarity in expression patterns between the DFCIX sequences in the test sample and the reference sample indicates the test cell sample contains cells from a tumor that is of the same stage as those tumor cells in the reference sample (for tumor-cell bearing reference samples). Conversely, a difference in expression patterns between the test cell sample and reference cell sample (for tumor-cell bearing control samples) indicates the test cell sample is not from the same tumor size as the tumor cell or cells in the reference cell sample.

Expression of DFCIs: 32-35 is decreased in tumors larger than 1.5 cm as compared to small tumors. Accordingly, a decrease in expression of one or more of DFCIs: 32-35 in a test cell population relative to reference cell population made up of non-cancerous cells indicates a large tumor is present in the subject.

In preferred embodiments, the expression of 2, 3, 4, 5, or more of the sequences represented by DFCIs: 32-35 are measured. If desired, expression of these sequences can be measured along with other sequences whose expression is known to be altered according to tumor size.

Diagnosing a neoplasm

The invention further provides a method of diagnosing a neoplasm, *e.g.*, a solid tumor such as a breast, lung, colon, prostate or stomach tumor in a subject. A neoplasm is diagnosed by examining the expression of one or more DFCIX nucleic acid sequences from a test population of cells that contain a suspected tumor. The population of cells may contain the

primary tumor, *e.g.*, breast tissue in the case of a breast tumor, or may alternatively contain cells into which a primary tumor has disseminated, *e.g.*, blood or lymphatic fluid. Preferably, the test cell population is a from blood.

Expression of one or more of the DFCIX nucleic acid sequences, *e.g.*, DFCIs: 1-52 is measured in the test cell and compared to the expression of the sequences in the reference cell population. Preferably, three or more of DFCIs: 11, 12, 45-51 and 52 nucleic acid sequences are measured. More preferably, expression of four or more DFCIs: , 26, 28, 33, 36-52 nucleic acids sequences are measured. More preferably, expression of four or more of DFCIs: 26, 28, 33, 36-43 and 44 nucleic acid sequences are measured. The reference cell population contains at least one cell whose neoplastic state is known. If the reference cell population contains no neoplastic cells, than a similarity in expression between DFCIX sequences in the test population and the reference cell population indicates the test cell population does not contain a neoplastic cell. A difference in expression between DFCIX sequences in the test population and the reference cell population indicates the reference cell population contains a neoplastic cell.

Conversely, when the reference cell population contains neoplastic cells, a similarity in expression pattern between the test cell population and the reference cell population indicates the test cell population includes a neoplastic cell. A difference in expression between DFCIX sequences in the test population and the reference cell population indicates the reference cell population contains a non-neoplastic cell.

Assessing efficacy of treatment of a neoplasm in a subject

The differentially expressed DFCIX sequences identified herein also allow for the course of treatment of a neoplasm to be monitored. In this method, a test cell population is provided from a subject undergoing treatment for a neoplasm. If desired, test cell populations can be taken from the subject at various time points before, during, or after treatment. Expression of one or more of the DFCIX sequences, *e.g.*, DFCIs: 1-52, in the cell population is then measured and compared to a reference cell population which includes cells whose neoplastic state is known. In an alternative embodiment four or more of DFCIs: 26-28, 33, 36-43 and 44 are measured. In another alternative embodiment three or more of DFCIs: 11, 12, 45-51 and 52 are measured. Preferably, the reference cells not been exposed to the treatment.

If the reference cell population contains no neoplastic cells, a similarity in expression between DFCIX sequences in the test cell population and the reference cell population indicates

that the treatment is efficacious. However, a difference in expression between DFCIX sequences in the test population and this reference cell population indicates the treatment is not efficacious.

By "efficacious" is meant that the treatment leads to a decrease in size or metastatic potential of a neoplasm in a subject, or a shift in tumor stage to a less advanced stage. When treatment is applied prophylactically, "efficacious" means that the treatment retards or prevents a neoplasm from forming.

When the reference cell population contains neoplastic cells, *e.g.*, when the reference cell population includes cancerous cells taken from the subject at the time of diagnosis but prior to beginning treatment, an alteration in the expression pattern between the test cell population and the reference cell population indicates the treatment is not efficacious. In contrast, a difference in expression between DFCIX sequences in the test population and this reference cell population indicates the treatment is efficacious.

When the reference cell population contains non-neoplastic cells, a decrease in expression of one or more of the sequences DFCIs: 29-31 or an increase in expression of one or more of the sequences DFCIs: 1-28 and 32-35 indicates the treatment efficacious.

Efficaciousness can be determined in association with any known method for treating a neoplasm. In some embodiments, the treatment is with an anti-estrogen agent, preferably tamoxifen.

Selecting a therapeutic agent for treating a neoplasm that is appropriate for a particular individual

Differences in the genetic makeup of individuals can result in differences in their relative abilities to metabolize various drugs. An agent that is metabolized in a subject to act as an anti-neoplastic agent can manifest itself by inducing a change in gene expression pattern in the subject's cells from that characteristic of a neoplastic state to a gene expression pattern characteristic of a non-neoplastic state. Accordingly, the differentially expressed DFCIX sequences disclosed herein allow for a putative therapeutic or prophylactic anti-neoplastic agent to be tested in a test cell population from a selected subject in order to determine if the agent is a suitable anti-neoplastic agent in the subject.

To identify an anti-neoplastic agent, that is appropriate for a specific subject, a test cell population from the subject is exposed to a therapeutic agent, and the expression of one or more of DFCIs: 1-52 sequences is measured.

The test cell population contains the primary tumor, *e.g.*, a breast carcinoma, or a bodily fluid, such as, *e.g.*, blood or lymph fluid, into which a tumor cell has disseminated. For example atest cell population is incubated in the presence of a candidate therapeutic agent and the pattern of gene expression of the test sample is measures and compared to one or more reference profiles, *e.g.*, a cancer reference profile or a non-cancer reference profile. Alternatively, the agent is first mixed with a cell extract, *e.g.*, a liver cell extract, which contains enzymes that metabolize drugs into an active form. The activated form of the therapeutic agent can then be mixed with the test cell population and gene expression measured. Preferably, the cell population is contacted *ex vivo* with the agent or activated form of the agent.

Expression of the nucleic acid sequences in the test cell population is then compared to the expression of the nucleic acid sequences a reference cell population. The reference cell population includes at least one cell whose neoplastic state is known. If the reference cell is non-cancerous, a similar gene expression profile between the test cell population and the reference cell population indicates the agent is suitable for treating the neoplasm in the subject. A difference in expression between sequences in the test cell population and those in the reference cell population indicates that the agent is not suitable for treating the neoplasm in the subject.

If the reference cell is cancerous, a similarity in gene expression patterns between the test cell population and the reference cell population indicates the agent is not suitable for treating the neoplasm in the subject, while similar gene expression patterns indicate the agent will be suitable for treating the subject.

In some embodiments, a decrease in expression of one or more of the sequences DFCIs: 29-31 or an increase in expression of one or more of the sequences DFCIs: 1- 28 and 32-35 in a test cell population relative to a reference cell population containing cancerous cells is indicative that the agent is therapeutic.

The test agent can be any compound or composition. In some embodiments the test agents are compounds and composition know to be anti-cancer agents, *e.g.*, an anti-estrogen agent such as tamoxifen.

Screening assays for identifying a candidate therapeutic agent for treating or preventing a neoplasm

The differentially expressed sequences disclosed herein can also be used to identify candidate therapeutic agents for treating a neoplasm. The method is based on screening a candidate therapeutic agent to determine if it converts an expression profile of DFCI: 1-52 sequences characteristic of a cancerous state to a pattern indicative of a noncancerous state.

In the method a cell is exposed to a test agent or a combination of test agents (sequentially or consequentially) and the expression of one or more DFCI: 1-52 sequences in the cell is measured. The expression of the DFCIX sequences in the test population is compared to expression level of the DFCIX sequences in a reference cell population that is not exposed to the test agent. Test agents will increase the expression of DFCIX sequences that are downregulated in some cancerous cells, and/or will decrease the expression of those DFCIX sequences that are upregulated in cancerous cells.

In some embodiments, the reference cell population includes cancerous cells. When this cell population is used, an alteration in expression of the nucleic acid sequences in the presence of the agent from the expression profile of the cell population in the absence of the agent indicates the agent is a candidate therapeutic agent for treating a neoplasm.

The test agent can be a compound not previously described or can be a previously known compound but which is not known to be an anti-neoplastic agent.

An agent effective in stimulating expression of underexpressed genes, or in suppressing expression of overexpressed genes can be further tested for its ability to prevent tumor growth, *e.g.*, carcinoma growth, and is a potential therapeutic useful for the treatment of such tumors. Further evaluation of the clinical usefulness of such a compound can be performed using standard methods of evaluating toxicity and clinical effectiveness of anti-cancer agents.

Categorizing a neoplasm in a subject

By comparing expression patterns of DFCIX sequences in test cell populations containing neoplastic cells with those in a reference cell population, neoplasms can be categorized in a subject.

The method includes providing a cell population containing at least one neoplastic cell from a subject and measuring the expression of one or more DFCIX nucleic acid sequences in

the cell. Expression of the nucleic acid sequences in the test cell population is compared to the expression of the nucleic acid sequences in a reference cell population comprising at least one cell whose neoplastic state and category is known. A similarity in expression patterns in the test cell population and the reference cell population indicates the cancerous cell in the test cell population is of same neoplastic state and category as that present in the reference cell population.

Assessing the prognosis of a subject with a neoplasm

Also provided is a method of assessing the prognosis of a subject containing a neoplasm by comparing the expression of one or more DFCIX sequences in a test cell population to the expression of the sequences in a reference cell population derived from patients over a spectrum of disease stages. By comparing gene expression of one or more DFCIX sequences in the test cell population and the reference cell population(s), or by comparing the pattern of gene expression overtime in test cell populations derived from the subject, the prognosis of the subject can be assessed.

The reference cell population includes primarily noncancerous or cancerous cells. When the reference cell population includes primarily noncancerous cells, which are often ER-negative a decrease in expression of one or more of the sequences DFCIs: 20-28 and 32-35, or an increase of expression of one or more of the sequences DFCIs: 29-31, indicates less favorable prognosis. An increase in expression of one or more of the sequences DFCIs: 1-19 indicates a more favorable prognosis. Conversely, an increase in expression of one or more of the sequences DFCIs: 20-28 and 32-35, or a decrease in expression of sequences DFCIs: 29-31 indicates a more favorable prognosis for the subject, while a decrease in expression of sequence DFCIs: 1-19 suggests a less favorable prognosis.

Alternatively, when a reference cell population includes primarily noncancerous cells, an increase in expression of one or more or the sequences DFCIs 11, 22, 26, 28, 33, 36-52 indicates a less favorable prognosis in the subject, while a decrease or similar expression indicates a more favorable prognosis.

The reference cell population can include primarily cancerous cells, which can be either ER-positive or ER-negative. When the reference cell population included primarily ER-positive cancerous cells, an increase in expression of one or more of the sequences DFCIs: 20-28 and 32-35, or a decrease in expression of one or more of the sequences DFCIs: 29-31, in the test cell

population compared to the reference cell population indicates a less favorable prognosis. Conversely, a decrease in expression of one or more of the sequences DFCIs: 20-31, or an increase in expression of sequences DFCIs: 1-19 and 32-35 indicates a more favorable prognosis for the subject.

Alternatively when a reference cell population includes primarily disseminated cancerous cells a decrease in expression of one or more or the sequences DFCIs 11, 22, 26, 28, 33, 36-52 indicates a more favorable prognosis in the subject, while a increase or similar expression indicates a more favorable prognosis.

Kits

The invention also includes a DFCIX-detection reagent, e.g., nucleic acids that specifically identify one or more DFCIX nucleic acids by having homologous nucleic acid sequences, such as oligonucleotide sequences, complementary to a portion of the DFCIX nucleic acids or antibodies to proteins encoded by the DFCIX nucleic acids packaged together in the form of a kit. The kit may contain in separate containers a nucleic acid or antibody (either already bound to a solid matrix or packaged separately with reagents for binding them to the matrix) , control formulations (positive and/or negative), and/or a detectable label. Instructions (e.g., written, tape, VCR, CD-ROM, etc.) for carrying out the assay may be included in the kit. The assay may for example be in the form of a Northern hybridization or a sandwich ELISA as known in the art.

For example, DFCIX detection reagent, is immobilized on a solid matrix such as a porous strip to form at least one DFCIX detection site. The measurement or detection region of the porous strip may include a plurality of sites containing a nucleic acid. A test strip may also contain sites for negative and/or positive controls. Alternatively, control sites are located on a separate strip from the test strip. Optionally, the different detection sites may contain different amounts of immobilized nucleic acids, *i.e.*, a higher amount in the first detection site and lesser amounts in subsequent sites. Upon the addition of test sample, the number of sites displaying a detectable signal provides a quantitative indication of the amount of DFCIX present in the sample. The detection sites may be configured in any suitably detectable shape and are typically in the shape of a bar or dot spanning the width of a teststrip.

Alternatively, the kit contains a nucleic acid substrate array comprising one or more nucleic acid sequences. The nucleic acids on the array specifically identify one or more nucleic

acid sequences represented by DFCIs: 1-52. In various embodiments, the expression of 2, 3,4, 5, 6, 7,8, 9, 10, 15, 20, 25, 40 or 50 or more of the sequences represented by DFCIs: 1-52 are identified by virtue if binding to the array. The substrate array can be on, *e.g.*, a solid substrate, *e.g.*, a "chip" as described in U.S. Patent No.5,744,305.

Arrays and pluralities

The invention also includes a nucleic acid substrate array comprising one or more nucleic acid sequences. The nucleic acids on the array specifically identify one or more nucleic acid sequences represented by DFCIs: 1-52. In various embodiments, the expression of 2, 3,4, 5, 6, 7,8, 9, 10, 15, 20, 25, 40 or 50 or more of the sequences represented by DFCIs: 1-52 are identified.

The nucleic acids in the array can identify the enumerated nucleic acids by, *e.g.*, having homologous nucleic acid sequences, such as oligonucleotide sequences, complementary to a portion of the recited nucleic acids. The substrate array can be on, *e.g.*, a solid substrate, *e.g.*, a "chip" as described in U.S. Patent No.5,744,305.

The invention also includes an isolated plurality (*i.e.*, a mixture if two or more nucleic acids) of nucleic acid sequences. The nucleic acid sequence can be in a liquid phase or a solid phase, *e.g.*, immobilized on a solid support such as a nitrocellulose membrane. The plurality typically includes one or more of the nucleic acid sequences represented by DFCIs: 1-52. In various embodiments, the plurality includes 2, 3,4, 5, 6, 7,8, 9, 10, 15, 20, 25, 40 or 50 or more of the sequences represented by DFCIs: 1-52.

An "isolated" nucleic acid molecule is one that is separated from other nucleic acid molecules, which are present in the natural source of the nucleic acid. Examples of isolated nucleic acid molecules include, but are not limited to, recombinant DNA molecules contained in a vector, recombinant DNA molecules maintained in a heterologous host cell, partially or substantially purified nucleic acid molecules, and synthetic DNA or RNA molecules. Preferably, an "isolated" nucleic acid is free of sequences which naturally flank the nucleic acid (*i.e.*, sequences located at the 5' and 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. In various embodiments, the isolated nucleic acid molecule can contain less than about 50 kb, 25 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of nucleotide sequences which naturally flank the nucleic acid molecule in genomic DNA of the

cell from which the nucleic acid is derived. Moreover, an "isolated" nucleic acid molecule, such as a cDNA molecule, can be substantially free of other cellular material or culture medium when produced by recombinant techniques, or of chemical precursors or other chemicals when chemically synthesized.

Methods of treating neoplasms

The invention provides a method for treating a neoplasm in a subject. The neoplasm can be, *e.g.*, a carcinoma, particularly a carcinoma of mammary epithelial tissue. Administration can be prophylactic or therapeutic to a subject at risk of (or susceptible to) a disorder or having a disorder associated with aberrant expression or activity of the herein described differentially expressed sequences (*e.g.*, DFCIs: 1-52).

The therapeutic method includes increasing the expression, or function, or both of one or more gene products of genes whose expression is decreased ("underexpressed genes") in a cancerous cell relative to normal cells of the tissue type from which the cancerous carcinoma cells are derived. In these methods, the subject is treated with an effective amount of a compound, which increases the amount of one of more of the underexpressed genes in the subject. Administration can be systemic or local, *e.g.*, in the immediate vicinity of, the subject's cancerous cells. This compound could be, for example, the polypeptide product of the underexpressed gene, or a biologically active fragment thereof. In addition, the compound could be a nucleic acid encoding the underexpressed gene and having expression control elements permitting expression in the carcinoma cells; or an agent which increases the level of expression of such gene endogenous to the carcinoma cells (i.e., which up-regulates expression of the underexpressed gene or genes). Treating a patient with one or more of the underexpressed gene product, or a biologically active fragment of the underexpressed gene product, will help counter the effects of down-regulation of the gene or genes in the subject's cancerous cells and improve the clinical condition of the subject

The method also includes decreasing the expression, or function, or both, of one or more gene products of genes whose expression is increased ("overexpressed gene") in a cancerous cell as compared to a non-cancerous cell. Expression can be inhibited in any of several ways known in the art. For example, expression can be inhibited by administering to the subject a nucleic acid that inhibits, or antagonizes, the expression of the overexpressed gene or genes. In one

embodiment, an antisense oligonucleotide can be administered which disrupts expression of the gene or genes.

Alternatively, function of one or more gene products of the overexpressed genes can be inhibited by administering a compound that binds to or otherwise inhibits the function of the gene products. The compound can be, *e.g.*, an antibody to the overexpressed gene product or gene products.

These modulatory methods can be performed *ex vivo* or *in vitro* (*e.g.*, by culturing the cell with the agent) or, alternatively, *in vivo* (*e.g.*, by administering the agent to a subject). As such, the present invention provides methods of treating an individual afflicted with a disease or disorder characterized by aberrant expression or activity of the differentially expressed proteins or nucleic acid molecules. In one embodiment, the method involves administering an agent (*e.g.*, an agent identified by a screening assay described herein), or combination of agents that modulates (*e.g.*, upregulates or downregulates) expression or activity of one or more differentially expressed genes. In another embodiment, the method involves administering a protein or combination of proteins or a nucleic acid molecule or combination of nucleic acid, molecules as therapy to compensate for reduced or aberrant expression or activity of the differentially expressed genes.

Diseases and disorders that are characterized by increased (relative to a subject not suffering from the disease or disorder) levels or biological activity of the genes may be treated with therapeutics that antagonize (*i.e.*, reduce or inhibit) activity of the overexpressed gene or genes. Therapeutics that antagonize activity may be administered therapeutically or prophylactically.

Therapeutics that may be utilized include, *e.g.*, (*i*) a polypeptide, or analogs, derivatives, fragments or homologs thereof of the overexpressed or underexpressed sequence or sequences; (*ii*) antibodies to the overexpressed or underexpressed sequence or sequences; (*iii*) nucleic acids encoding the over or underexpressed sequence or sequences; (*iv*) antisense nucleic acids or nucleic acids that are "dysfunctional" (*i.e.*, due to a heterologous insertion within the coding sequences of coding sequences of one or more overexpressed or underexpressed sequences); or (*v*) modulators (*i.e.*, inhibitors, agonists and antagonists that alter the interaction between an over/underexpressed polypeptide and its binding partner. The dysfunctional antisense moleculed

are utilized to "knockout" endogenous function of a polypeptide by homologous recombination (see, *e.g.,* Capecchi, *Science* 244: 1288-1292 1989)

Diseases and disorders that are characterized by decreased (relative to a subject not suffering from the disease or disorder) levels or biological activity may be treated with therapeutics that increase (*i.e.*, are agonists to) activity. Therapeutics that upregulate activity may be administered in a therapeutic or prophylactic manner. Therapeutics that may be utilized include, but are not limited to, a polypeptide (or analogs, derivatives, fragments or homologs thereof) or an agonist that increases bioavailability.

Increased or decreased levels can be readily detected by quantifying peptide and/or RNA, by obtaining a patient tissue sample (*e.g.,* from biopsy tissue) and assaying it *in vitro* for RNA or peptide levels, structure and/or activity of the expressed peptides (or mRNAs of a gene whose expression is altered). Methods that are well-known within the art include, but are not limited to, immunoassays (*e.g.,* by Western blot analysis, immunoprecipitation followed by sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis, immunocytochemistry, etc.) and/or hybridization assays to detect expression of mRNAs (*e.g.,* Northern assays, dot blots, *in situ* hybridization, etc.).

Administration of a prophylactic agent can occur prior to the manifestation of symptoms characteristic of aberrant gene expression, such that a disease or disorder is prevented or, alternatively, delayed in its progression. Depending on the type of aberrant expression detected, the agent can be used for treating the subject. The appropriate agent can be determined based on screening assays described herein.

Another aspect of the invention pertains to methods of modulating expression or activity of one of the herein described differentially regulated genes for therapeutic purposes. The method includes contacting a cell with an agent that modulates one or more of the activities of the gene products of the differentially expressed genes. An agent that modulates protein activity can be an agent as described herein, such as a nucleic acid or a protein, a naturally-occurring cognate ligand of these proteins, a peptide, a peptidomimetic, or other small molecule. In one embodiment, the agent stimulates one or more protein activities of one or more of the differentially expressed genes. Examples of such stimulatory agents include active protein and a nucleic acid molecule encoding such proteins that has been introduced into the cell.

Pharmaceutical compositions for treating neoplasms

In another aspect the invention includes pharmaceutical, or therapeutic, compositions containing one or more therapeutic compounds described herein. Pharmaceutical formulations may include those suitable for oral, rectal, nasal, topical (including buccal and sub-lingual), vaginal or parenteral (including intramuscular, sub-cutaneous and intravenous) administration, or for administration by inhalation or insufflation. The formulations may, where appropriate, be conveniently presented in discrete dosage units and may be prepared by any of the methods well known in the art of pharmacy. All such pharmacy methods include the steps of bringing into association the active compound with liquid carriers or finely divided solid carriers or both as needed and then, if necessary, shaping the product into the desired formulation.

Pharmaceutical formulations suitable for oral administration may conveniently be presented as discrete units, such as capsules, cachets or tablets, each containing a predetermined amount of the active ingredient; as a powder or granules; or as a solution, a suspension or as an emulsion. The active ingredient may also be presented as a bolus electuary or paste, and be in a pure form, i.e., without a carrier. Tablets and capsules for oral administration may contain conventional excipients such as binding agents, fillers, lubricants, disintegrant or wetting agents. A tablet may be made by compression or molding, optionally with one or more formulational ingredients. Compressed tablets may be prepared by compressing in a suitable machine the active ingredients in a free-flowing form such as a powder or granules, optionally mixed with a binder, lubricant, inert diluent, lubricating, surface active or dispersing agent. Molded tablets may be made by molding in a suitable machine a mixture of the powdered compound moistened with an inert liquid diluent. The tablets may be coated according to methods well known in the art. Oral fluid preparations may be in the form of, for example, aqueous or oily suspensions, solutions, emulsions, syrups or elixirs, or may be presented as a dry product for constitution with water or other suitable vehicle before use. Such liquid preparations may contain conventional additives such as suspending agents, emulsifying agents, non-aqueous vehicles (which may include edible oils), or preservatives. The tablets may optionally be formulated so as to provide slow or controlled release of the active ingredient therein.

Formulations for parenteral administration include aqueous and non-aqueous sterile injection solutions which may contain anti-oxidants, buffers, bacteriostats and solutes which render the formulation isotonic with the blood of the intended recipient; and aqueous and non-aqueous sterile suspensions which may include suspending agents and thickening agents. The formulations may be presented in unit dose or multi-dose containers, for example sealed

ampoules and vials, and may be stored in a freeze-dried (lyophilized) condition requiring only the addition of the sterile liquid carrier, for example, saline, water-for-injection, immediately prior to use. Alternatively, the formulations may be presented for continuous infusion. Extemporaneous injection solutions and suspensions may be prepared from sterile powders, granules and tablets of the kind previously described.

Formulations for rectal administration may be presented as a suppository with the usual carriers such as cocoa butter or polyethylene glycol. Formulations for topical administration in the mouth, for example buccally or sublingually, include lozenges, comprising the active ingredient in a flavored base such as sucrose and acacia or tragacanth, and pastilles comprising the active ingredient in a base such as gelatin and glycerin or sucrose and acacia. For intra-nasal administration the compounds of the invention may be used as a liquid spray or dispersible powder or in the form of drops. Drops may be formulated with an aqueous or non-aqueous base also comprising one or more dispersing agents, solubilizing agents or suspending agents. Liquid sprays are conveniently delivered from pressurized packs.

For administration by inhalation the compounds are conveniently delivered from an insufflator, nebulizer, pressurized packs or other convenient means of delivering an aerosol spray. Pressurized packs may comprise a suitable propellant such as dichlorodifluoromethane, trichlorofluoromethane, dichiorotetrafluoroethane, carbon dioxide or other suitable gas. In the case of a pressurized aerosol, the dosage unit may be determined by providing a valve to deliver a metered amount.

Alternatively, for administration by inhalation or insufflation, the compounds may take the form of a dry powder composition, for example a powder mix of the compound and a suitable powder base such as lactose or starch. The powder composition may be presented in unit dosage form, in for example, capsules, cartridges, gelatin or blister packs from which the powder may be administered with the aid of an inhalator or insuffiator.

When desired, the above described formulations, adapted to give sustained release of the active ingredient, may be employed. The pharmaceutical compositions may also contain other active ingredients such as antimicrobial agents, immunosuppressants or preservatives.

It should be understood that in addition to the ingredients particularly mentioned above, the formulations of this invention may include other agents conventional in the art having regard

to the type of formulation in question, for example, those suitable for oral administration may include flavoring agents.

Preferred unit dosage formulations are those containing an effective dose, as recited below, or an appropriate fraction thereof, of the active ingredient.

For each of the aforementioned conditions, the compositions may be administered orally or via injection at a dose of from about 0.1 to about 250 mg/kg per day. The dose range for adult humans is generally from about 5 mg to about 17.5 g/day, preferably about 5 mg to about 10 g/day, and most preferably about 100 mg to about 3 g/day. Tablets or other unit dosage forms of presentation provided in discrete units may conveniently contain an amount which is effective at such dosage or as a multiple of the same, for instance, units containing about 5 mg to about 500 mg, usually from about 100 mg to about 500 mg.

The pharmaceutical composition preferably is administered orally or by injection (intravenous or subcutaneous), and the precise amount administered to a subject will be the responsibility of the attendant physician. However, the dose employed will depend upon a number of factors, including the age and sex of the subject, the precise disorder being treated, and its severity. Also the route of administration may vary depending upon the condition and its severity.

Novel nucleic acids whose expression is differentially regulated in tumors and non-tumor tissues

The invention also provides composition of novel nucleic acid sequences that are differentially regulated in tumors and non-tumor tissues. Accordingly, the invention includes isolated nucleic acid molecules, which includes the nucleotide sequence of either DFCIs: 7, 8,25, or31 or a fragment of one or more of these sequences.

DFCI 7

A DFCI 7 nucleic acid includes the sequence:

```
1     TGCCGAAGCT TTTTTTTTTT AACAGTACAA TGATACAGAG AATTTATTCA ATTCATACAA
61    GTAATTTACC AGATCTAAAC AGTGAATAGA CTGACCTAAG GGGAAAAAAA TCCTTACATT
121   GATAGCAAAA TATCTTCTGG CCCCCATAAA TAATCTGCAA TCATTTGCTA GGAAAAAAAC
181   TTCATAACCA TATGGGTCAT GCAGACATTT TTATTTATTT (SEQ ID NO:1)
```

DFCI 8

A DFCI 8 nucleic acid includes the sequence:

```
1   TGCCGAAKCT TGAGCCTGAC TGTGCATCTC TAGGMTTAGA AYAWTAAWWW GAAKCACTTT
61  AGAAAAMAMA KCTTGTGGGA AAKCCTAAHT CKGCTCATAT GCTYAAACTG GACKGCGGCT
121 GAACGTGCKA TGAGAKGGCW CGATCATYCT ACTRTKGGCG CACCTKAMTA TGHAMTGGRH
181 CATGCTTHTH TAH (SEQ ID NO:2)
```

DFCI 25

A DFCI 25 nucleic acid includes the sequence:

```
1     TGCGACCTAT TATTCGTCTC ACATGGAGAA TACAAAGTAT GAGAAGTAAA AAATGAGCCC
61    ATATTTTCTC ACTTAGTGGT GAGGGGAGAA GAGTAAGTGG ATAGAGGGCA AAGGGAAGCA
121   GTCAGGTAAC TATATCCCAA TTTTGTTACA GGCTCCAATT TGGAGTTTCA GAGGTTTGTA
181   GATAGTTTTA ATATTACCCT GAGGGAGGGA AGTCTTCTAG CCAGAGTCTT GTTTTAGAGT
241   GTAGAAAGGA ACCAGTCTTT CCTGCATACA CTCCGGCATG CA (SEQ ID NO:3)
```

DFCI 31

A DFCI 31 nucleic acid includes the sequence:

```
1     GAGTGATGST TTTTTCAAAT TSTNTTATGA AACNNNAAAT GTCTATTCCT NNNTTTCCGG
61    GTGTGGTAGA AGAATATGAA AAGATCAAAA GTGGGTGACT TCCAGNGTAA CAATTT (SEQ ID
NO:4)
```

The nucleic acids of the invention include those that encode a DFCIX polypeptide or protein. As used herein, the terms polypeptide and protein are interchangeable.

In some embodiments, a DFCIX nucleic acid encodes a mature DFCIX polypeptide. As used herein, a "mature" form of a polypeptide or protein described herein relates to the product of a naturally occurring polypeptide or precursor form or proprotein. The naturally occurring polypeptide, precursor or proprotein includes, by way of nonlimiting example, the full-length gene product, encoded by the corresponding gene. Alternatively, it may be defined as the polypeptide, precursor or proprotein encoded by an open reading frame described herein. The product "mature" form arises, again by way of nonlimiting example, as a result of one or more naturally occurring processing steps that may take place within the cell in which the gene product arises. Examples of such processing steps leading to a "mature" form of a polypeptide or protein include the cleavage of the N-terminal methionine residue encoded by the initiation codon of an open reading frame, or the proteolytic cleavage of a signal peptide or leader sequence. Thus a mature form arising from a precursor polypeptide or protein that has residues 1 to N, where residue 1 is the N-terminal methionine, would have residues 2 through N remaining after removal of the N-terminal methionine. Alternatively, a mature form arising from a precursor

polypeptide or protein having residues 1 to N, in which an N-terminal signal sequence from residue 1 to residue M is cleaved, would have the residues from residue M+1 to residue N remaining. Further as used herein, a "mature" form of a polypeptide or protein may arise from a step of post-translational modification other than a proteolytic cleavage event. Such additional processes include, by way of non-limiting example, glycosylation, myristoylation or phosphorylation. In general, a mature polypeptide or protein may result from the operation of only one of these processes, or a combination of any of them.

Among the DFCIX nucleic acids is the nucleic acid whose sequence is provided in SEQ ID NO: 7, 8, 25 or 31, or a fragment thereof. Additionally, the invention includes mutant or variant nucleic acids of SEQ ID NO: 7, 8, 25 or 31, or a fragment thereof, any of whose bases may be changed from the corresponding bases shown in SEQ ID NO: 7, 8, 25 or 31, while still encoding a protein that maintains at least one of its DFCIX-like activities and physiological functions (*i.e.*, modulating angiogenesis, neuronal development). The invention further includes the complement of the nucleic acid sequence of SEQ ID NO: 7, 8, 25 or 31, including fragments, derivatives, analogs and homologs thereof. The invention additionally includes nucleic acids or nucleic acid fragments, or complements thereto, whose structures include chemical modifications.

One aspect of the invention pertains to isolated nucleic acid molecules that encode DFCIX proteins or biologically active portions thereof. Also included are nucleic acid fragments sufficient for use as hybridization probes to identify DFCIX-encoding nucleic acids (*e.g.*, DFCIX mRNA) and fragments for use as polymerase chain reaction (PCR) primers for the amplification or mutation of DFCIX nucleic acid molecules. As used herein, the term "nucleic acid molecule" is intended to include DNA molecules (*e.g.*, cDNA or genomic DNA), RNA molecules (*e.g.*, mRNA), analogs of the DNA or RNA generated using nucleotide analogs, and derivatives, fragments and homologs thereof. The nucleic acid molecule can be single-stranded or double-stranded, but preferably is double-stranded DNA.

"Probes" refer to nucleic acid sequences of variable length, preferably between at least about 10 nucleotides (nt), 100 nt, or as many as about, *e.g.*, 6,000 nt, depending on use. Probes are used in the detection of identical, similar, or complementary nucleic acid sequences. Longer length probes are usually obtained from a natural or recombinant source, are highly specific and much slower to hybridize than oligomers. Probes may be single- or double-stranded and designed to have specificity in PCR, membrane-based hybridization technologies, or ELISA-like technologies.

An "isolated" nucleic acid molecule is one that is separated from other nucleic acid

molecules that are present in the natural source of the nucleic acid. Examples of isolated nucleic acid molecules include, but are not limited to, recombinant DNA molecules contained in a vector, recombinant DNA molecules maintained in a heterologous host cell, partially or substantially purified nucleic acid molecules, and synthetic DNA or RNA molecules. Preferably, an "isolated" nucleic acid is free of sequences which naturally flank the nucleic acid (*i.e.*, sequences located at the 5' and 3' ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. For example, in various embodiments, the isolated DFCIX nucleic acid molecule can contain less than about 50 kb, 25 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of nucleotide sequences which naturally flank the nucleic acid molecule in genomic DNA of the cell from which the nucleic acid is derived. Moreover, an "isolated" nucleic acid molecule, such as a cDNA molecule, can be substantially free of other cellular material or culture medium when produced by recombinant techniques, or of chemical precursors or other chemicals when chemically synthesized.

A nucleic acid molecule of the present invention, *e.g.*, a nucleic acid molecule having the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31, or a complement of any of this nucleotide sequence, can be isolated using standard molecular biology techniques and the sequence information provided herein. Using all or a portion of the nucleic acid sequence of SEQ ID NO: 7, 8, 25 or 31, as a hybridization probe, DFCIX nucleic acid sequences can be isolated using standard hybridization and cloning techniques (*e.g.*, as described in Sambrook *et al.*, eds., MOLECULAR CLONING: A LABORATORY MANUAL 2nd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989; and Ausubel, *et al.*, eds., CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, New York, NY, 1993.)

A nucleic acid of the invention can be amplified using cDNA, mRNA or alternatively, genomic DNA, as a template and appropriate oligonucleotide primers according to standard PCR amplification techniques. The nucleic acid so amplified can be cloned into an appropriate vector and characterized by DNA sequence analysis. Furthermore, oligonucleotides corresponding to DFCIX nucleotide sequences can be prepared by standard synthetic techniques, *e.g.*, using an automated DNA synthesizer.

As used herein, the term "oligonucleotide" refers to a series of linked nucleotide residues, which oligonucleotide has a sufficient number of nucleotide bases to be used in a PCR reaction. A short oligonucleotide sequence may be based on, or designed from, a genomic or cDNA sequence and is used to amplify, confirm, or reveal the presence of an identical, similar or complementary DNA or RNA in a particular cell or tissue. Oligonucleotides comprise portions of a nucleic acid sequence having about 10 nt, 50 nt, or 100 nt in length, preferably about 15 nt to

30 nt in length. In one embodiment, an oligonucleotide comprising a nucleic acid molecule less than 100 nt in length would further comprise at lease 6 contiguous nucleotides of SEQ ID NO: 7, 8, 25 or 31, or a complement thereof. Oligonucleotides may be chemically synthesized and may be used as probes.

In another embodiment, an isolated nucleic acid molecule of the invention comprises a nucleic acid molecule that is a complement of the nucleotide sequence shown in SEQ ID NO: 7, 8, 25 or 31, or a portion of this nucleotide sequence. A nucleic acid molecule that is complementary to the nucleotide sequence shown in SEQ ID NO: 7, 8, 25 or 31 is one that is sufficiently complementary to the nucleotide sequence shown in SEQ ID NO: 7, 8, 25 or 31 that it can hydrogen bond with little or no mismatches to the nucleotide sequence shown in SEQ ID NO: 7, 8, 25 or 31, thereby forming a stable duplex.

As used herein, the term "complementary" refers to Watson-Crick or Hoogsteen base pairing between nucleotide units of a nucleic acid molecule, and the term "binding" means the physical or chemical interaction between two polypeptides or compounds or associated polypeptides or compounds or combinations thereof. Binding includes ionic, non-ionic, Von der Waals, hydrophobic interactions, etc. A physical interaction can be either direct or indirect. Indirect interactions may be through or due to the effects of another polypeptide or compound. Direct binding refers to interactions that do not take place through, or due to, the effect of another polypeptide or compound, but instead are without other substantial chemical intermediates.

Moreover, the nucleic acid molecule of the invention can comprise only a portion of the nucleic acid sequence of SEQ ID NO: 7, 8, 25 or 31, *e.g.*, a fragment that can be used as a probe or primer, or a fragment encoding a biologically active portion of DFCIX. Fragments provided herein are defined as sequences of at least 6 (contiguous) nucleic acids or at least 4 (contiguous) amino acids, a length sufficient to allow for specific hybridization in the case of nucleic acids or for specific recognition of an epitope in the case of amino acids, respectively, and are at most some portion less than a full length sequence. Fragments may be derived from any contiguous portion of a nucleic acid or amino acid sequence of choice. Derivatives are nucleic acid sequences or amino acid sequences formed from the native compounds either directly or by modification or partial substitution. Analogs are nucleic acid sequences or amino acid sequences that have a structure similar to, but not identical to, the native compound but differs from it in respect to certain components or side chains. Analogs may be synthetic or from a different evolutionary origin and may have a similar or opposite metabolic activity compared to wild type.

Derivatives and analogs may be full length or other than full length, if the derivative or analog contains a modified nucleic acid or amino acid, as described below. Derivatives or

analogs of the nucleic acids or proteins of the invention include, but are not limited to, molecules comprising regions that are substantially homologous to the nucleic acids or proteins of the invention, in various embodiments, by at least about 70%, 80%, 85%, 90%, 95%, 98%, or even 99% identity (with a preferred identity of 80-99%) over a nucleic acid or amino acid sequence of identical size or when compared to an aligned sequence in which the alignment is done by a computer homology program known in the art, or whose encoding nucleic acid is capable of hybridizing to the complement of a sequence encoding the aforementioned proteins under stringent, moderately stringent, or low stringent conditions. See *e.g.* Ausubel, *et al.*, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, New York, NY, 1993, and below. An exemplary program is the Gap program (Wisconsin Sequence Analysis Package, Version 8 for UNIX, Genetics Computer Group, University Research Park, Madison, WI) using the default settings, which uses the algorithm of Smith and Waterman (Adv. Appl. Math., 1981, 2: 482-489, which is incorporated herein by reference in its entirety).

A "homologous nucleic acid sequence" or "homologous amino acid sequence," or variations thereof, refer to sequences characterized by a homology at the nucleotide level or amino acid level as discussed above. Homologous nucleotide sequences encode those sequences coding for isoforms of a DFCIX polypeptide. Isoforms can be expressed in different tissues of the same organism as a result of, for example, alternative splicing of RNA. Alternatively, isoforms can be encoded by different genes. In the present invention, homologous nucleotide sequences include nucleotide sequences encoding for a DFCIX polypeptide of species other than humans, including, but not limited to, mammals, and thus can include, *e.g.*, mouse, rat, rabbit, dog, cat cow, horse, and other organisms. Homologous nucleotide sequences also include, but are not limited to, naturally occurring allelic variations and mutations of the nucleotide sequences set forth herein. A homologous nucleotide sequence does not, however, include the nucleotide sequence encoding human DFCIX protein. Homologous nucleic acid sequences include those nucleic acid sequences that encode conservative amino acid substitutions (see below), as well as a polypeptide having DFCIX activity. Biological activities of the DFCIX proteins are described below. A homologous amino acid sequence does not encode the amino acid sequence of a human DFCIX polypeptide.

The nucleotide sequence determined from the cloning of the human DFCIX gene allows for the generation of probes and primers designed for use in identifying and/or cloning DFCIX homologues in other cell types, *e.g.*, from other tissues, as well as DFCIX homologues from other mammals. The probe/primer typically comprises a substantially purified oligonucleotide. The oligonucleotide typically comprises a region of nucleotide sequence that hybridizes under

stringent conditions to at least about 12, 25, 50, 100, 150, 200, 250, 300, 350 or 400 or more consecutive sense strand nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31; or an anti-sense strand nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31; or of a naturally occurring mutant of SEQ ID NO: 7, 8, 25 or 31.

Probes based on the human DFCIX nucleotide sequence can be used to detect transcripts or genomic sequences encoding the same or homologous proteins. In various embodiments, the probe further comprises a label group attached thereto, *e.g.*, the label group can be a radioisotope, a fluorescent compound, an enzyme, or an enzyme co-factor. Such probes can be used as a part of a diagnostic test kit for identifying cells or tissue which misexpress a DFCIX protein, such as by measuring a level of a DFCIX-encoding nucleic acid in a sample of cells from a subject *e.g.*, detecting DFCIX mRNA levels or determining whether a genomic DFCIX gene has been mutated or deleted.

A "polypeptide having a biologically active portion of DFCIX" refers to polypeptides exhibiting activity similar, but not necessarily identical to, an activity of a polypeptide of the present invention, including mature forms, as measured in a particular biological assay, with or without dose dependency. A nucleic acid fragment encoding a "biologically active portion of DFCIX" can be prepared by isolating a portion of SEQ ID NO: 7, 8, 25 or 31 that encodes a polypeptide having a DFCIX biological activity (biological activities of the DFCIX proteins are described below), expressing the encoded portion of DFCIX protein (*e.g.*, by recombinant expression *in vitro*) and assessing the activity of the encoded portion of DFCIX. For example, a nucleic acid fragment encoding a biologically active portion of DFCIX can optionally include an ATP-binding domain. In another embodiment, a nucleic acid fragment encoding a biologically active portion of DFCIX includes one or more regions.

DFCIX Variants

The invention further encompasses nucleic acid molecules that differ from the nucleotide sequences shown in SEQ ID NO: 7, 8, 25 or 31 due to the degeneracy of the genetic code. These nucleic acids thus encode the same DFCIX protein as that encoded by the nucleotide sequence shown in SEQ ID NO: 7, 8, 25 or 31.

In addition to the human DFCIX nucleotide sequence shown in SEQ ID NO: 7, 8, 25 or 31, it will be appreciated by those skilled in the art that DNA sequence polymorphisms that lead to changes in the amino acid sequences of DFCIX may exist within a population (*e.g.*, the human population). Such genetic polymorphism in the DFCIX gene may exist among individuals within a population due to natural allelic variation. As used herein, the terms "gene" and

"recombinant gene" refer to nucleic acid molecules comprising an open reading frame encoding a DFCIX protein, preferably a mammalian DFCIX protein. Such natural allelic variations can typically result in 1-5% variance in the nucleotide sequence of the DFCIX gene. Any and all such nucleotide variations and resulting amino acid polymorphisms in DFCIX that are the result of natural allelic variation and that do not alter the functional activity of DFCIX are intended to be within the scope of the invention.

Moreover, nucleic acid molecules encoding DFCIX proteins from other species, and thus that have a nucleotide sequence that differs from the human sequence of SEQ ID NO: 7, 8, 25 or 31 are intended to be within the scope of the invention. Nucleic acid molecules corresponding to natural allelic variants and homologues of the DFCIX cDNAs of the invention can be isolated based on their homology to the human DFCIX nucleic acids disclosed herein using the human cDNAs, or a portion thereof, as a hybridization probe according to standard hybridization techniques under stringent hybridization conditions. For example, a soluble human DFCIX cDNA can be isolated based on its homology to human membrane-bound DFCIX. Likewise, a membrane-bound human DFCIX cDNA can be isolated based on its homology to soluble human DFCIX.

Accordingly, in another embodiment, an isolated nucleic acid molecule of the invention is at least 6 nucleotides in length and hybridizes under stringent conditions to the nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31. In another embodiment, the nucleic acid is at least 10, 25, 50, 100, 250, 500 or 750 nucleotides in length. In another embodiment, an isolated nucleic acid molecule of the invention hybridizes to the coding region. As used herein, the term "hybridizes under stringent conditions" is intended to describe conditions for hybridization and washing under which nucleotide sequences at least 60% homologous to each other typically remain hybridized to each other.

Homologs (*i.e.*, nucleic acids encoding DFCIX proteins derived from species other than human) or other related sequences (*e.g.*, paralogs) can be obtained by low, moderate or high stringency hybridization with all or a portion of the particular human sequence as a probe using methods well known in the art for nucleic acid hybridization and cloning.

As used herein, the phrase "stringent hybridization conditions" refers to conditions under which a probe, primer or oligonucleotide will hybridize to its target sequence, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures than shorter sequences. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. The Tm is the

temperature (under defined ionic strength, pH and nucleic acid concentration) at which 50% of the probes complementary to the target sequence hybridize to the target sequence at equilibrium. Since the target sequences are generally present at excess, at Tm, 50% of the probes are occupied at equilibrium. Typically, stringent conditions will be those in which the salt concentration is less than about 1.0 M sodium ion, typically about 0.01 to 1.0 M sodium ion (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes, primers or oligonucleotides (*e.g.*, 10 nt to 50 nt) and at least about 60°C for longer probes, primers and oligonucleotides. Stringent conditions may also be achieved with the addition of destabilizing agents, such as formamide.

Stringent conditions are known to those skilled in the art and can be found in CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, N.Y. (1989), 6.3.1-6.3.6. Preferably, the conditions are such that sequences at least about 65%, 70%, 75%, 85%, 90%, 95%, 98%, or 99% homologous to each other typically remain hybridized to each other. A non-limiting example of stringent hybridization conditions is hybridization in a high salt buffer comprising 6X SSC, 50 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA, and 500 mg/ml denatured salmon sperm DNA at 65°C. This hybridization is followed by one or more washes in 0.2X SSC, 0.01% BSA at 50°C. An isolated nucleic acid molecule of the invention that hybridizes under stringent conditions to the sequence of SEQ ID NO: 7, 8, 25 or 31 corresponds to a naturally occurring nucleic acid molecule. As used herein, a "naturally-occurring" nucleic acid molecule refers to an RNA or DNA molecule having a nucleotide sequence that occurs in nature (*e.g.*, encodes a natural protein).

In a second embodiment, a nucleic acid sequence that is hybridizable to the nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31, or fragments, analogs or derivatives thereof, under conditions of moderate stringency is provided. A non-limiting example of moderate stringency hybridization conditions are hybridization in 6X SSC, 5X Denhardt's solution, 0.5% SDS and 100 mg/ml denatured salmon sperm DNA at 55°C, followed by one or more washes in 1X SSC, 0.1% SDS at 37°C. Other conditions of moderate stringency that may be used are well known in the art. See, *e.g.*, Ausubel *et al.* (eds.), 1993, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, NY, and Kriegler, 1990, GENE TRANSFER AND EXPRESSION, A LABORATORY MANUAL, Stockton Press, NY.

In a third embodiment, a nucleic acid that is hybridizable to the nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31, or fragments, analogs or derivatives thereof, under conditions of low stringency, is provided. A non-limiting example of low stringency hybridization conditions are hybridization in 35% formamide, 5X SSC, 50 mM

Tris-HCl (pH 7.5), 5 mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.2% BSA, 100 mg/ml denatured salmon sperm DNA, 10% (wt/vol) dextran sulfate at 40°C, followed by one or more washes in 2X SSC, 25 mM Tris-HCl (pH 7.4), 5 mM EDTA, and 0.1% SDS at 50°C. Other conditions of low stringency that may be used are well known in the art (*e.g.*, as employed for cross-species hybridizations). See, *e.g.*, Ausubel *et al.* (eds.), 1993, CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, NY, and Kriegler, 1990, GENE TRANSFER AND EXPRESSION, A LABORATORY MANUAL, Stockton Press, NY; Shilo and Weinberg, 1981, *Proc Natl Acad Sci USA 78*: 6789-6792.

Conservative mutations

In addition to naturally-occurring allelic variants of the DFCIX sequence that may exist in the population, the skilled artisan will further appreciate that changes can be introduced by mutation into the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31, thereby leading to changes in the amino acid sequence of the encoded DFCIX protein, without altering the functional ability of the DFCIX protein. For example, nucleotide substitutions leading to amino acid substitutions at "non-essential" amino acid residues can be made in the sequence of SEQ ID NO: 7, 8, 25 or 31. A "non-essential" amino acid residue is a residue that can be altered from the wild-type sequence of DFCIX without altering the biological activity, whereas an "essential" amino acid residue is required for biological activity. For example, amino acid residues that are conserved among the DFCIX proteins of the present invention, are predicted to be particularly unamenable to alteration.

Another aspect of the invention pertains to nucleic acid molecules encoding DFCIX proteins that contain changes in amino acid residues that are not essential for activity. An isolated nucleic acid molecule encoding a DFCIX protein homologous to the protein of can be created by introducing one or more nucleotide substitutions, additions or deletions into the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31, such that one or more amino acid substitutions, additions or deletions are introduced into the encoded protein.

Mutations can be introduced into the nucleotide sequence of SEQ ID NO: 7, 8, 25 or 31 by standard techniques, such as site-directed mutagenesis and PCR-mediated mutagenesis. Preferably, conservative amino acid substitutions are made at one or more predicted non-essential amino acid residues. A "conservative amino acid substitution" is one in which the amino acid residue is replaced with an amino acid residue having a similar side chain. Families of amino acid residues having similar side chains have been defined in the art. These families include amino acids with basic side chains (*e.g.*, lysine, arginine, histidine), acidic side chains (*e.g.*, aspartic acid, glutamic acid), uncharged polar side chains (*e.g.*, glycine, asparagine,

glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (*e.g.*, alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side chains (*e.g.*, threonine, valine, isoleucine) and aromatic side chains (*e.g.*, tyrosine, phenylalanine, tryptophan, histidine). Thus, a predicted nonessential amino acid residue in DFCIX is replaced with another amino acid residue from the same side chain family. Alternatively, in another embodiment, mutations can be introduced randomly along all or part of a DFCIX coding sequence, such as by saturation mutagenesis, and the resultant mutants can be screened for DFCIX biological activity to identify mutants that retain activity. Following mutagenesis of SEQ ID NO: 7, 8, 25 or 31 the encoded protein can be expressed by any recombinant technology known in the art and the activity of the protein can be determined.

In one embodiment, a mutant DFCIX protein can be assayed for (1) the ability to form protein:protein interactions with other DFCIX proteins, other cell-surface proteins, or biologically active portions thereof, (2) complex formation between a mutant DFCIX protein and a DFCIX receptor; (3) the ability of a mutant DFCIX protein to bind to an intracellular target protein or biologically active portion thereof; (*e.g.*, avidin proteins); (4) the ability to bind DFCIX protein; or (5) the ability to specifically bind an anti-DFCIX protein antibody.

Antisense DFCIX Nucleic Acids

Another aspect of the invention pertains to isolated antisense nucleic acid molecules that are hybridizable to or complementary to the nucleic acid molecule comprising the nucleotide sequence of SEQ ID NO: 7,8, 25 or 31, or fragments, analogs or derivatives thereof. An "antisense" nucleic acid comprises a nucleotide sequence that is complementary to a "sense" nucleic acid encoding a protein, *e.g.*, complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. In specific aspects, antisense nucleic acid molecules are provided that comprise a sequence complementary to at least about 10, 25, 50, 100, 250 or 500 nucleotides or an entire DFCIX coding strand, or to only a portion thereof. Nucleic acid molecules encoding fragments, homologs, derivatives and analogs of a DFCIX protein or antisense nucleic acids complementary to a DFCIX nucleic acid sequence of SEQ ID NO: 7, 8, 25 or 31 are additionally provided.

In one embodiment, an antisense nucleic acid molecule is antisense to a "coding region" of the coding strand of a nucleotide sequence encoding DFCIX. The term "coding region" refers to the region of the nucleotide sequence comprising codons which are translated into amino acid residues. In another embodiment, the antisense nucleic acid molecule is antisense to a "noncoding region" of the coding strand of a nucleotide sequence encoding DFCIX. The term

"noncoding region" refers to 5' and 3' sequences which flank the coding region that are not translated into amino acids (*i.e.*, also referred to as 5' and 3' untranslated regions).

Given the coding strand sequences encoding DFCIX disclosed herein (*e.g.*, SEQ ID NO: 7, 8, 25 or 31), antisense nucleic acids of the invention can be designed according to the rules of Watson and Crick or Hoogsteen base pairing. The antisense nucleic acid molecule can be complementary to the entire coding region of DFCIX mRNA, but more preferably is an oligonucleotide that is antisense to only a portion of the coding or noncoding region of DFCIX mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of DFCIX mRNA. An antisense oligonucleotide can be, for example, about 5, 10, 15, 20, 25, 30, 35, 40, 45 or 50 nucleotides in length. An antisense nucleic acid of the invention can be constructed using chemical synthesis or enzymatic ligation reactions using procedures known in the art. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used.

Examples of modified nucleotides that can be used to generate the antisense nucleic acid include: 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid (v), 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, (acp3)w, and 2,6-diaminopurine. Alternatively, the antisense nucleic acid can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (*i.e.*, RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest, described further in the following subsection).

The antisense nucleic acid molecules of the invention are typically administered to a subject or generated *in situ* such that they hybridize with or bind to cellular mRNA and/or

genomic DNA encoding a DFCIX protein to thereby inhibit expression of the protein, *e.g.*, by inhibiting transcription and/or translation. The hybridization can be by conventional nucleotide complementarity to form a stable duplex, or, for example, in the case of an antisense nucleic acid molecule that binds to DNA duplexes, through specific interactions in the major groove of the double helix. An example of a route of administration of antisense nucleic acid molecules of the invention includes direct injection at a tissue site. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then administered systemically. For example, for systemic administration, antisense molecules can be modified such that they specifically bind to receptors or antigens expressed on a selected cell surface, *e.g.*, by linking the antisense nucleic acid molecules to peptides or antibodies that bind to cell surface receptors or antigens. The antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. To achieve sufficient intracellular concentrations of antisense molecules, vector constructs in which the antisense nucleic acid molecule is placed under the control of a strong pol II or pol III promoter are preferred.

In yet another embodiment, the antisense nucleic acid molecule of the invention is an α-anomeric nucleic acid molecule. An α-anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual β-units, the strands run parallel to each other (Gaultier *et al.* (1987) *Nucleic Acids Res* 15: 6625-6641). The antisense nucleic acid molecule can also comprise a 2'-o-methylribonucleotide (Inoue *et al.* (1987) *Nucleic Acids Res* 15: 6131-6148) or a chimeric RNA -DNA analogue (Inoue *et al.* (1987) *FEBS Lett* 215: 327-330).

Such modifications include, by way of nonlimiting example, modified bases, and nucleic acids whose sugar phosphate backbones are modified or derivatized. These modifications are carried out at least in part to enhance the chemical stability of the modified nucleic acid, such that they may be used, for example, as antisense binding nucleic acids in therapeutic applications in a subject.

DFCIX Ribozymes and PNA moieties

In still another embodiment, an antisense nucleic acid of the invention is a ribozyme. Ribozymes are catalytic RNA molecules with ribonuclease activity that are capable of cleaving a single-stranded nucleic acid, such as a mRNA, to which they have a complementary region. Thus, ribozymes (*e.g.*, hammerhead ribozymes (described in Haselhoff and Gerlach (1988) *Nature* 334:585-591)) can be used to catalytically cleave DFCIX mRNA transcripts to thereby inhibit translation of DFCIX mRNA. A ribozyme having specificity for a DFCIX-encoding

nucleic acid can be designed based upon the nucleotide sequence of a DFCIX DNA disclosed herein (*i.e.*, SEQ ID NO: 7, 8, 25 or 31). For example, a derivative of a Tetrahymena L-19 IVS RNA can be constructed in which the nucleotide sequence of the active site is complementary to the nucleotide sequence to be cleaved in a DFCIX-encoding mRNA. See, *e.g.*, Cech *et al.* U.S. Pat. No. 4,987,071; and Cech *et al.* U.S. Pat. No. 5,116,742. Alternatively, DFCIX mRNA can be used to select a catalytic RNA having a specific ribonuclease activity from a pool of RNA molecules. See, *e.g.*, Bartel *et al.*, (1993) *Science* 261:1411-1418.

Alternatively, DFCIX gene expression can be inhibited by targeting nucleotide sequences complementary to the regulatory region of the DFCIX (*e.g.*, the DFCIX promoter and/or enhancers) to form triple helical structures that prevent transcription of the DFCIX gene in target cells. See generally, Helene. (1991) *Anticancer Drug Des.* 6: 569-84; Helene. *et al.* (1992) *Ann. N.Y. Acad. Sci.* 660:27-36; and Maher (1992) *Bioassays* 14: 807-15.

In various embodiments, the nucleic acids of DFCIX can be modified at the base moiety, sugar moiety or phosphate backbone to improve, *e.g.*, the stability, hybridization, or solubility of the molecule. For example, the deoxyribose phosphate backbone of the nucleic acids can be modified to generate peptide nucleic acids (see Hyrup *et al.* (1996) *Bioorg Med Chem* 4: 5-23). As used herein, the terms "peptide nucleic acids" or "PNAs" refer to nucleic acid mimics, *e.g.*, DNA mimics, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone and only the four natural nucleobases are retained. The neutral backbone of PNAs has been shown to allow for specific hybridization to DNA and RNA under conditions of low ionic strength. The synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described in Hyrup *et al.* (1996) above; Perry-O'Keefe *et al.* (1996) *PNAS* 93: 14670-675.

PNAs of DFCIX can be used in therapeutic and diagnostic applications. For example, PNAs can be used as antisense or antigene agents for sequence-specific modulation of gene expression by, *e.g.*, inducing transcription or translation arrest or inhibiting replication. PNAs of DFCIX can also be used, *e.g.*, in the analysis of single base pair mutations in a gene by, *e.g.*, PNA directed PCR clamping; as artificial restriction enzymes when used in combination with other enzymes, *e.g.*, S1 nucleases (Hyrup B. (1996) above); or as probes or primers for DNA sequence and hybridization (Hyrup *et al.* (1996), above; Perry-O'Keefe (1996), above).

In another embodiment, PNAs of DFCIX can be modified, *e.g.*, to enhance their stability or cellular uptake, by attaching lipophilic or other helper groups to PNA, by the formation of PNA-DNA chimeras, or by the use of liposomes or other techniques of drug delivery known in the art. For example, PNA-DNA chimeras of DFCIX can be generated that may combine the

advantageous properties of PNA and DNA. Such chimeras allow DNA recognition enzymes, *e.g.*, RNase H and DNA polymerases, to interact with the DNA portion while the PNA portion would provide high binding affinity and specificity. PNA-DNA chimeras can be linked using linkers of appropriate lengths selected in terms of base stacking, number of bonds between the nucleobases, and orientation (Hyrup (1996) above). The synthesis of PNA-DNA chimeras can be performed as described in Hyrup (1996) above and Finn *et al.* (1996) *Nucl Acids Res* 24: 3357-63. For example, a DNA chain can be synthesized on a solid support using standard phosphoramidite coupling chemistry, and modified nucleoside analogs, *e.g.*, 5'-(4-methoxytrityl) amino-5'-deoxy-thymidine phosphoramidite, can be used between the PNA and the 5' end of DNA (Mag *et al.* (1989) *Nucl Acid Res* 17: 5973-88). PNA monomers are then coupled in a stepwise manner to produce a chimeric molecule with a 5' PNA segment and a 3' DNA segment (Finn *et al.* (1996) above). Alternatively, chimeric molecules can be synthesized with a 5' DNA segment and a 3' PNA segment. See, Petersen *et al.* (1975) *Bioorg Med Chem Lett* 5: 1119-11124.

In other embodiments, the oligonucleotide may include other appended groups such as peptides (*e.g.*, for targeting host cell receptors *in vivo*), or agents facilitating transport across the cell membrane (see, *e.g.*, Letsinger *et al.*, 1989, *Proc. Natl. Acad. Sci. U.S.A.* 86:6553-6556; Lemaitre *et al.*, 1987, Proc. *Natl. Acad. Sci.* 84:648-652; PCT Publication No. W088/09810) or the blood-brain barrier (see, *e.g.*, PCT Publication No. W089/10134). In addition, oligonucleotides can be modified with hybridization triggered cleavage agents (See, *e.g.*, Krol *et al.*, 1988, *BioTechniques* 6:958-976) or intercalating agents. (See, *e.g.*, Zon, 1988. *Pharm. Res.* 5: 539-549). To this end, the oligonucleotide may be conjugated to another molecule, *e.g.*, a peptide, a hybridization triggered cross-linking agent, a transport agent, a hybridization-triggered cleavage agent, etc.

DFCIX Polypeptides

The invention also provided DFCIX polypeptides encoded by DFCIX nucleic acids. In some embodiments, up to 20% or more of the residues may be so changed in the mutant or variant protein. In some embodiments, the DFCIX polypeptide according to the invention is a mature polypeptide.

In general, a DFCIX -like variant that preserves DFCIX-like function includes any variant in which residues at a particular position in the sequence have been substituted by other amino acids, and further include the possibility of inserting an additional residue or residues between two residues of the parent protein as well as the possibility of deleting one or more residues from the parent sequence. Any amino acid substitution, insertion, or deletion is

encompassed by the invention. In favorable circumstances, the substitution is a conservative substitution as defined above.

One aspect of the invention pertains to isolated DFCIX proteins, and biologically active portions thereof, or derivatives, fragments, analogs or homologs thereof. Also provided are polypeptide fragments suitable for use as immunogens to raise anti-DFCIX antibodies. In one embodiment, native DFCIX proteins can be isolated from cells or tissue sources by an appropriate purification scheme using standard protein purification techniques. In another embodiment, DFCIX proteins are produced by recombinant DNA techniques. Alternative to recombinant expression, a DFCIX protein or polypeptide can be synthesized chemically using standard peptide synthesis techniques.

An "isolated" or "purified" protein or biologically active portion thereof is substantially free of cellular material or other contaminating proteins from the cell or tissue source from which the DFCIX protein is derived, or substantially free from chemical precursors or other chemicals when chemically synthesized. The language "substantially free of cellular material" includes preparations of DFCIX protein in which the protein is separated from cellular components of the cells from which it is isolated or recombinantly produced. In one embodiment, the language "substantially free of cellular material" includes preparations of DFCIX protein having less than about 30% (by dry weight) of non-DFCIX protein (also referred to herein as a "contaminating protein"), more preferably less than about 20% of non-DFCIX protein, still more preferably less than about 10% of non-DFCIX protein, and most preferably less than about 5% non-DFCIX protein. When the DFCIX protein or biologically active portion thereof is recombinantly produced, it is also preferably substantially free of culture medium, *i.e.*, culture medium represents less than about 20%, more preferably less than about 10%, and most preferably less than about 5% of the volume of the protein preparation.

The language "substantially free of chemical precursors or other chemicals" includes preparations of DFCIX protein in which the protein is separated from chemical precursors or other chemicals that are involved in the synthesis of the protein. In one embodiment, the language "substantially free of chemical precursors or other chemicals" includes preparations of DFCIX protein having less than about 30% (by dry weight) of chemical precursors or non-DFCIX chemicals, more preferably less than about 20% chemical precursors or non-DFCIX chemicals, still more preferably less than about 10% chemical precursors or non-DFCIX chemicals, and most preferably less than about 5% chemical precursors or non-DFCIX chemicals.

Biologically active portions of a DFCIX protein include peptides comprising amino acid

sequences sufficiently homologous to or derived from the amino acid sequence of the DFCIX protein, *e.g.*, the amino acid sequence shown in SEQ ID NO: 2, 4, 6, 8, 10, 12, 14 or 16 that include fewer amino acids than the full length DFCIX proteins, and exhibit at least one activity of a DFCIX protein. Typically, biologically active portions comprise a domain or motif with at least one activity of the DFCIX protein. A biologically active portion of a DFCIX protein can be a polypeptide which is, for example, 10, 25, 50, 100 or more amino acids in length. A biologically active portion of a DFCIX protein of the present invention may contain at least one of the above-identified domains conserved between the DFCIX proteins, *e.g.* TSR modules. Moreover, other biologically active portions, in which other regions of the protein are deleted, can be prepared by recombinant techniques and evaluated for one or more of the functional activities of a native DFCIX protein.

Determining homology between two or more sequence

To determine the percent homology of two amino acid sequences or of two nucleic acids, the sequences are aligned for optimal comparison purposes (*e.g.*, gaps can be introduced in either of the sequences being compared for optimal alignment between the sequences). The amino acid residues or nucleotides at corresponding amino acid positions or nucleotide positions are then compared. When a position in the first sequence is occupied by the same amino acid residue or nucleotide as the corresponding position in the second sequence, then the molecules are homologous at that position (*i.e.*, as used herein amino acid or nucleic acid "homology" is equivalent to amino acid or nucleic acid "identity").

The nucleic acid sequence homology may be determined as the degree of identity between two sequences. The homology may be determined using computer programs known in the art, such as GAP software provided in the GCG program package. See, *Needleman and Wunsch* 1970 *J Mol Biol* 48: 443-453. Using GCG GAP software with the following settings for nucleic acid sequence comparison: GAP creation penalty of 5.0 and GAP extension penalty of 0.3, the coding region of the analogous nucleic acid sequences referred to above exhibits a degree of identity preferably of at least 70%, 75%, 80%, 85%, 90%, 95%, 98%, or 99%, with the CDS (encoding) part of the DNA sequence shown in SEQ ID NO: 7, 8, 25 or 31.

The term "sequence identity" refers to the degree to which two polynucleotide or polypeptide sequences are identical on a residue-by-residue basis over a particular region of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over that region of comparison, determining the number of positions at which the identical nucleic acid base (*e.g.*, A, T, C, G, U, or I, in the case of nucleic acids) occurs in both sequences to yield the number of matched positions, dividing the number of

matched positions by the total number of positions in the region of comparison (*i.e.*, the window size), and multiplying the result by 100 to yield the percentage of sequence identity. The term "substantial identity" as used herein denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence that has at least 80 percent sequence identity, preferably at least 85 percent identity and often 90 to 95 percent sequence identity, more usually at least 99 percent sequence identity as compared to a reference sequence over a comparison region. The term "percentage of positive residues" is calculated by comparing two optimally aligned sequences over that region of comparison, determining the number of positions at which the identical and conservative amino acid substitutions, as defined above, occur in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the region of comparison (*i.e.*, the window size), and multiplying the result by 100 to yield the percentage of positive residues.

Chimeric and fusion proteins

The invention also provides DFCIX chimeric or fusion proteins. As used herein, a DFCIX "chimeric protein" or "fusion protein" comprises a DFCIX polypeptide operatively linked to a non-DFCIX polypeptide. An "DFCIX polypeptide" refers to a polypeptide having an amino acid sequence corresponding to DFCIX, whereas a "non-DFCIX polypeptide" refers to a polypeptide having an amino acid sequence corresponding to a protein that is not substantially homologous to the DFCIX protein, *e.g.*, a protein that is different from the DFCIX protein and that is derived from the same or a different organism. Within a DFCIX fusion protein the DFCIX polypeptide can correspond to all or a portion of a DFCIX protein. In one embodiment, a DFCIX fusion protein comprises at least one biologically active portion of a DFCIX protein. In another embodiment, a DFCIX fusion protein comprises at least two biologically active portions of a DFCIX protein. Within the fusion protein, the term "operatively linked" is intended to indicate that the DFCIX polypeptide and the non-DFCIX polypeptide are fused in-frame to each other. The non-DFCIX polypeptide can be fused to the N-terminus or C-terminus of the DFCIX polypeptide.

For example, in one embodiment a DFCIX fusion protein comprises a DFCIX polypeptide operably linked to the extracellular domain of a second protein. Such fusion proteins can be further utilized in screening assays for compounds that modulate DFCIX activity (such assays are described in detail below).

In another embodiment, the fusion protein is a GST-DFCIX fusion protein in which the DFCIX sequences are fused to the C-terminus of the GST (i.e., glutathione S-transferase)

sequences. Such fusion proteins can facilitate the purification of recombinant DFCIX.

In another embodiment, the fusion protein is a DFCIX-immunoglobulin fusion protein in which the DFCIX sequences comprising one or more domains are fused to sequences derived from a member of the immunoglobulin protein family. The DFCIX-immunoglobulin fusion proteins of the invention can be incorporated into pharmaceutical compositions and administered to a subject to inhibit an interaction between a DFCIX ligand and a DFCIX protein on the surface of a cell, to thereby suppress DFCIX-mediated signal transduction *in vivo*. In one nonlimiting example, a contemplated DFCIX ligand of the invention is the DFCIX receptor. The DFCIX-immunoglobulin fusion proteins can be used to affect the bioavailability of a DFCIX cognate ligand. Inhibition of the DFCIX ligand/DFCIX interaction may be useful therapeutically for both the treatment of proliferative and differentiative disorders, *e,g.,* . cancer as well as modulating (*e.g.*, promoting or inhibiting) cell survival, as well as acute and chronic inflammatory disorders and hyperplastic wound healing, e.g. hypertrophic scars and keloids. Moreover, the DFCIX-immunoglobulin fusion proteins of the invention can be used as immunogens to produce anti-DFCIX antibodies in a subject, to purify DFCIX ligands, and in screening assays to identify molecules that inhibit the interaction of DFCIX with a DFCIX ligand.

A DFCIX chimeric or fusion protein of the invention can be produced by standard recombinant DNA techniques. For example, DNA fragments coding for the different polypeptide sequences are ligated together in-frame in accordance with conventional techniques, *e.g.*, by employing blunt-ended or stagger-ended termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and enzymatic ligation. In another embodiment, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers that give rise to complementary overhangs between two consecutive gene fragments that can subsequently be annealed and reamplified to generate a chimeric gene sequence (see, for example, Ausubel et al. (eds.) CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, 1992). Moreover, many expression vectors are commercially available that already encode a fusion moiety (*e.g.*, a GST polypeptide). A DFCIX-encoding nucleic acid can be cloned into such an expression vector such that the fusion moiety is linked in-frame to the DFCIX protein.

The invention will be further described in the following examples, which do not limit the scope of the invention described in the claims. The following examples illustrate the identification and characterization of genes differentially expressed in normal and cancerous breast cells.

EXAMPLE 1: COMPARISON OF GENES DIFFERENTIALLY EXPRESSED IN CANCEROUS AND NON-CANCEROUS BREAST CELLS USING DIFFERENTIAL DISPLAY ANALYSIS

To identify clinically relevant gene expression changes associated with breast cancer, a differential display (DD) analysis was used as a first screen.

The DD method was used to compare normal breast myoepithelial, luminal epithelial (Clarke et al., Epithelial Cell. Biol., 3: 38-46, 1994), and 76N cells (Band et al., Proc. Natl. Acad. Sci. U.S.A., 86: 1249-53, 1989) with a highly malignant breast tumor cell line, MDA-MB-435. Cailleau et al., In Vitro, 14: 911-915, 1978. Normal breast myoepithelial and luminal epithelial cells were sorted from primary cultures of mammoplasty tissue by immunomagnetic methods using epithelial membrane antigen (EMA) and common acute lymphoblastic leukemia antigen (CALLA) antibodies. Clarke et al., Epithelial Cell. Biol., 3: 38-46, 1994. The 21T breast tumor progression series of cell lines originated from a primary tumor and metastatic pleural effusion from a single patient. Band et al., Cancer Res., 50: 7351-7357, 1990. 76N normal breast epithelial cells obtained from mammoplasty tissue express several myoepithelial cell markers and undergo senescence after 20 passages. Band et al., Proc. Natl. Acad. Sci. U.S.A., 86: 1249-1253, 1989. Other cell lines were obtained from the ATCC. All cells were grown in DFCI-1 medium (Band et al., Proc. Natl. Acad. Sci. U.S.A., 86: 1249-1253, 1989) and harvested at 70% confluence. The PT series of breast tumor tissue samples were obtained within 30 minutes of surgery and frozen in liquid nitrogen, while the H series was OCT-cryofrozen archived tumor tissue. All RNA was prepared by the CsCl-cushion method as described. Chirgwin, et al., Biochemistry, 18: 5294-5299, 1979; Martin et al., Methods Enzymol., 303: 234-258, 1999. ER status was determined clinically by the cytosolic ligand-based assay (H series) or immunoassay (PT series).

Differential display was performed as described (Martin et al., Methods Enzymol., 303: 234-258, 1999) to compare normal breast epithelial cells and a metastatic breast tumor cell line, MDA-MB-435. Cailleau et al., In Vitro, 14: 911-915, 1978. The normal cells used were either

76N cultured breast epithelial cells (Band et al., Proc. Natl: Acad. Sci. U.S.A., 86: 1249-1253, 1989) or sorted normal breast myoepithelial and luminal epithelial cells. Clarke et al., Epithelial Cell. Biol., 3: 38-46, 1994. Both up- and down-regulated cDNA bands were selected for analysis. Approximately 70 primer pairs were used, including (i) LHA-1, -2, -3, -4, -5, -6, -7 in combination with LHT_{11}-G, -A, -C (Martin et al., Methods Enzymol., 303: 234-258, 1999), (ii) E1-OPA-1, -2, -3, -7, -8, H3-OPA-4, -5, -6, -9, -10 in combination with LHT_{11}-G, -A, -C (Martin et al., Methods Enzymol., 303: 234-258, 1999), (iii) ARP-1, -2, -3, -4 in combination with AP-1, -2, -3, -4, -5, -6, -7, -8, -9, 10 (Genomyx/Beckman Corp., Foster City, CA), and (iv) AP-1, -2, -3, -4, -5 in combination with T12M-G, -A, -T, -C (GenHunter Corp., Nashville, TN).

PCR conditions were as recommended by the respective primer kit manufacturers or as described (Martin et al., Methods Enzymol., 303: 234-258, 1999) for the LHA and OPA series of primers. Gel electrophoresis was performed on the extended format of the programmable Genomyx LR apparatus (Genomyx/Beckman Corp., Foster City, CA). Differential display bands were eluted by boiling in ddH_2O and precipitated. Martin et al., Methods Mol. Biol., 85: 77-85, 1997. Either of three different approaches was then used to identify the bands and obtain a cDNA clone. (I) cDNAs were TA-cloned into pCR2 or pCR2.1 (Invitrogen, Carlesbad, CA) directly and colonies screened for differential expressors, which were then sequenced. (II) cDNAs were directly sequenced (Martin et al., Methods Mol. Biol., 85: 77-85, 1997; Martin et al., Biotechniques, 24: 1018-1026, 1998), a gene-specific primer was made and the PCR product was then TA-cloned and sequenced to confirm cloning of the correct cDNA. (III) cDNAs were directly sequenced, TA-cloned, and sequenced to confirm cloning of the correct cDNA. Genes were identified by querying GenBank using the BLAST algorithm (Altschul et al., Nucleic Acids Res., 25: 3389-3402, 1997) as described (Martin et al., Methods Mol. Biol., 85: 77-85, 1997).

Seventy primer combinations were used to screen approximately 7000 genes, which, since cells express 15,000 mRNAs (Alberts et al., Molecular Biology of the Cell, Garland Publishing, Inc., New York, 1994), represents over one-third of all expressed genes. 4.5% of all mRNAs were differential, hence, a total of 700 genes were differentially expressed between the cancer and normal cells. This number is in agreement with other studies using different methods. Velculescu et al., Science, 270: 484-487, 1995.). A set of 170 genes that were differentially expressed was identified. The 170 genes represent approximately one-quarter of all the genes that were differentially expressed in the normal and cancer cells compared, while the 107 genes included on the hybridization arrays represent one-seventh.

The great majority of the differential genes observed were similarly expressed in all of the normal breast cell types, but were either down- or up-regulated in the tumor cells. A marked contrast was seen in the types of genes that comprised the down- and up-regulated categories. Nearly 75% of the genes that were down-regulated in the tumor cells were categorized as filamentous, cell surface, and secreted genes that play roles in adhesion, communication, and the maintenance of cell shape. In contrast, 75% of the known genes whose expression was up-regulated in tumor cells were enzymes involved in metabolism, macromolecular synthesis, and disruption of the extracellular matrix.

EXAMPLE 2: COMPARISON OF GENES DIFFERENTIALLY EXPRESSED IN CANCEROUS AND NON-CANCEROUS BREAST CELLS USING HYBRIDIZATION ARRAY ANALYSIS

Messenger RNA expression patterns were further characterized using a membrane-based hybridization array spotted with tags representing 124 different genes. The gene tags included 89 DD-identified normal cell-specific genes, 18 DD-identified tumor cell-specific genes, as well as literature reported cancer genes and housekeeping genes.

Membrane arrays with tags for 124 different genes were made by spotting PCR products or whole plasmids using a hand-held 96-pin spotting devise. The templates for PCR were DD-isolated cDNA fragments that were either cloned into the pCR2.1 TA-cloning vector (Invitrogen, Carlesbad, CA) or were used directly following DD gel-elution without cloning. Cloned tags were amplified using M13 vector primers. Uncloned tags were amplified using one DD primer in combination with one gene-specific primer. Martin et al., Methods Mol. Biol., 85: 77-85, 1997; Martin et al. Biotechniques, 24: 1018-1026, 1998. After PCR, samples were purified on QIAquick PCR purification spin columns (Qiagen Inc., Valencia, CA). An aliquot of each was electrophoresed on an agarose gel and products were quantitated and size-checked by comparison to standards. In a few cases, whole plasmids consisting of cDNA inserted into the pCR2.1 TA-cloning vector (Invitrogen, Carlsbad, CA) or the vector alone were prepared using a mini-prep kit (Qiagen Inc., Valencia, CA), quantitated by electrophoresis, and arrayed directly. cDNA-tags were spotted onto positively-charged nylon membranes (Micron Separations Inc., Westboro, MA) using a multiprint 96-pin replicator with 16 offset positions (V&P Scientific, Inc., San Diego, CA) to give 1536 spots per 3.5 x 5 inch membrane. Each tag at a concentration of approximately 0.1 μg/μL was incubated 10 minutes in 0.4 M NaOH, 10 mM EDTA to denature,

then chilled on ice. An aliquot was diluted 1:30 and the tags were applied in quadruplicate at both concentrations, the stock 0.1 μg/μL and the diluted 0.003 μg/μL. The spotting devise delivered approximately 0.1 μL (as per manufacturer) with high reproducibility. In a control experiment where three different ^{32}P-labeled gene tags were applied to a membrane, which was then UV cross-linked, rinsed, and quantitated by phosphorimaging, the standard error of the mean (SEM) was 4-8% of the mean for 12 sets of 16 spots. Twenty replicate membranes with the 124 different gene tags were prepared in parallel, with thorough cleaning of the replicator pins between every three membranes. cDNA tags were UV cross-linked to the membranes then stored in sealed bags at 4°C.

Radiolabeled cDNA probes were prepared from 5 μg total cellular RNA by incorporating 50 μCi α^{32}P-dCTP into first strand cDNA as described. Martin et al., Methods Mol. Biol., 85: 77-85, 1997. Incorporation rates of 5 to 20% were standard. Probes with incorporation rates of less than 3% were not used. Membranes were pre-hybridized approximately 3 hours in a formamide-based hybridization buffer (ExpressHyb, Clonetech Corp., Palo Alto, CA) at 68°C. The entire radiolabeled cDNA probe was then added to the buffer and membranes were hybridized 14-18 hours at 68°C. Membranes were washed as recommended by the manufacturer of the hybridization solution and exposed to a phosphor-imaging screen for two days. After scanning, membranes were stripped and used again for a total of four hybridizations. MCF-10A and 21PT profiles were averaged from three repeated experiments each, while myoepithelial, luminal epithelial, MDA-MB-435, and PT-4 profiles were averaged from two repeated experiments each. Other profiles shown represent individual experiments.

To quantify signal intensities of the hybridized spots, equal-sized ellipses were drawn around all spots using software (ImageQuant) provided with the phosphorimager (Molecular Dynamics, Sunnyvale, CA). Data from only the higher concentration spots were used. Median background was subtracted and signals that were less than five-fold above background level were considered too low to accurately measure ("BKG"). Mean signals were calculated from quadruplicate measurable spots, or if three of the four spots were measurable. Sets with standard errors exceeding 150% of the mean were disregarded ("ND"). Signal intensities for each membrane were normalized to the median signal of that membrane. For RNAs that were run multiple times, geometric means of all non-BKG membrane-normalized values were calculated. A single median BKG value was determined from an entire set of membranes being compared and this value was substituted for all BKG values. Signals for each individual gene were then normalized to the geometric mean of the expression level of that gene across the set of

membranes being compared. Genes with consistently low signals across an entire set of comparison membranes were omitted from the analysis.

To assess the reproducibility of this set of hybridization array assays, experiments using the same RNA preparation were repeated on different days. Measurable, median-normalized expression values for each gene were then compared. In two experiments using MCF-10A RNA, expression levels of 95.5% of genes had repeated values that were within 4-fold of each other. In five experiments performed using MCF-10A RNA, 87% of gene expression values deviated from their respective means by less than 3-fold and 95% of values deviated by less than 5-fold . Based on the level of reproducibility of this set of array experiments, individual gene expression changes of less than 5-fold were not considered significant. In addition, no conclusions based on a single data point or an individual gene have been made. All conclusions presented are based on expression changes of greater than 5-fold that occurred in clusters comprised of at least 3 different genes.

The array assay as performed was highly sensitive for a non-PCR-based assay as indicated by the dynamic range, which covered nearly 4 orders of magnitude. Further, expression signals from more than 90% of the DD genes could be measured, which was considerably better than our detection rate of approximately 60% with conventional Northern assays for the same set of DD genes.

EXAMPLE 3: IDENTIFICATION OF GENE CLUSTERS WITH CLINICALLY RELEVANT EXPRESSION PATTERNS IN CANCEROUS AND NON-CANCEROUS BREAST CELLS

Messenger RNA expression levels of 124 genes were assayed using hybridization arrays in breast tumor tissue samples obtained from 18 patients and 7 breast cell lines.

Cluster analysis was used to organize the genes and the tissues so that those with the most closely related expression patterns were positioned adjacent to each other. Eisen et al., Proc. Natl. Acad. Sci. U.S.A., 95: 14863-14868, 1998. Cluster analysis is a computer method that calculates correlation coefficients between all gene and tissue data sets. It then organizes the positions of the rows and columns of the display and generates hierarchical trees that indicate the degree of relatedness by the height of the nodes.

When cluster analysis was performed using expression patterns from all of the DD genes included on the arrays, the cells and tumor tissues fell into two major groups. One group

included predominantly ER-positive samples, while a second group included predominantly ER-negative samples. This division indicated that the panel included a number of genes with expression patterns that reflected ER status.

Seven well-characterized breast cell lines were included in the analyzed tissues. The manner in which cluster analysis organized these cell lines demonstrated its ability to accurately recognize and make groupings according to physiologically relevant characteristics. The 21T series of cell lines (21MT-1, 21MT-2, 21NT, and 21PT), which were all derived from a single breast cancer patient (Band et al., Cancer Res., 50: 7351-7357, 1990), were grouped together in a closely related cluster. 21PT cells, which were unable to form tumors in nude mice (Band et al. Cancer Res., 50: 7351-7357, 1990), were positioned on a deeper node reflecting a more distant relationship to the three other 21T cell lines, which were all tumorigenic in mice. Band et al., Cancer Res., 50: 7351-7357, 1990. The highly metastatic breast tumor cell line MDA-MB-435 (MDA-435) (Cailleau et al., In Vitro, 14: 911-915, 1978) was found to be more similar to the 21T series of tumor cell lines than was the immortal but non-malignant cell line MCF-10A. Soule et al., Cancer Res., 50: 6075-6086, 1990. The only ER-positive cell line tested, MCF7 (Soule et al., Natl. Cancer Inst., 51: 1409-1416, 1973), was widely separated from the six ER-negative cell lines.

To identify genes with clinically relevant expression patterns, charts with tissues ordered by various clinical parameters were prepared. These were then screened for gene clusters with expression patterns that increased or decreased across the chart. The parameters screened included ER status, tumor stage, grade, size, the percentage of S-phase cells, and patient age. Gene clusters with mean expression levels that showed a statistically significant association with the clinical parameter were identified. Two clusters were significantly associated with ER status, one with clinical stage, and one with tumor size. No clusters were significantly associated with tumor grade, percentage of S-phase cells, or patient age.

Identities of genes in the two strongly ER-associated gene clusters are shown in Table 1. These clusters were expressed inversely to each other ($R=-0.50$, $p=0.012$) and expression of both was strongly correlated with ER status (I: $p=0.0002$, II: $p=0.0010$, Fisher's exact test). Expression of the p53 cluster (cluster I) was higher in ER-positive tissues and lower in ER-negative tissues. Expression of the maspin cluster (cluster II) was the inverse.

The normal breast myoepithelial and luminal epithelial cells used for DD comparisons were categorized as ER-positive by array analysis. The tumor cell line used for DD, MDA-MB-435, was ER-negative. This result is consistent with the isolation of many ER status-associated genes by DD.

A second major clinical grouping applied to breast cancer is tumor stage, which takes into account information on tumor size, nodal status, and distant metastases. Fisher et al., W. Cancer Medicine, 3rd edition, 2: 1706-1750, 1993. Gene cluster III, which included HSP-90 and two unknown genes, was significantly over-expressed in stage IV tumors relative to stage I, II, and III tumors ($p=0.0025$, Fisher's exact test). Stage IV breast tumors are distinguished from earlier stage tumors by the presence of distant metastases. Clinical stage is currently the best indicator of disease prognosis (Fisher et al., W. Cancer Medicine, 3rd edition, 2: 1706-1750, 1993), and hence this cluster may represent a valuable set of markers that provide prognostic information.

Tumor size is also an important independent predictor of disease prognosis (Fisher et al., W. Cancer Medicine, 3rd edition, 2: 1706-1750, 1993). Gene cluster IV, which included keratin 14, was reduced in expression in tumors larger than 1.5 cm relative to smaller tumors ($p=0.0406$, Fisher's exact test).

EXAMPLE 4: CLASSIFICATION OF BREAST TUMORS USING GENE EXPRESSION PATTERNS

Cluster analysis and the expression patterns of the four clinically relevant gene clusters were used to group breast tumor tissue into categories. Gene expression profiles and the contributions of the four selected clusters were determined, as well as tabulated clinical data for the 18 breast cancer patients.

Cluster analysis was performed using publicly available software written by M. Eisen, Stanford University (http://rana.stanford. edu/clustering). Data sets were logarithmically transformed and the similarity metric was an uncentered correlation. An image contrast of 3 or 4 was used.

Cluster analysis with the four selected gene clusters sorted the tumors into two major groups that differed in their ER status ($p=0.0002$). Grouping by other clinical parameters is also apparent. For example, the two highly related groups of tumors, group 1: H16, H4, H43, and group 2: PT-10, PT-6, included tumors with highly similar clinical data. The former group were advanced stage, ER-negative tumors, while the latter were ER-positive, stage II invasive carcinomas from women of similar ages.

Though cluster analysis was able to group breast tumors by their ER status with high accuracy, those grouped apparently inappropriately may represent the most interesting cases. Three of the 18 tumors appeared to be grouped inappropriately. A single ER-positive tumor (H16) grouped in the ER-negative cluster, and two ER-negative tumors (H10 and H33) grouped

in the ER-positive cluster. The gene expression patterns may reflect clinically important characteristics of these tumors. For instance, tumor H16 may express a receptor capable of binding to estrogen but otherwise non-functional and hence unable to activate ER responsive genes. Such a tumor would be predicted to be unresponsive to treatment with anti-estrogens. We note that H16 was an unusual tumor specimen. It was derived from a hip metastasis that appeared to originate from a primary breast tumor surgically removed 18 years prior. Following total hip replacement, the patient was treated with radiation and tamoxifen. After two years, the patient is without evidence of recurrent cancer. It is not possible to make conclusions regarding the tumor's responsiveness to tamoxifen, since tamoxifen was not the sole treatment. The other two inappropriately grouped tumors, H10 and H33, tumors may harbor constitutive oncogenic mutations, either in the ER itself or in downstream components of the pathway, that result in ER pathway activation in the absence of ER ligand-binding activity. Tumors of this type have been previously reported. Biswas et al., Mol. Med., 4: 454-467, 1998.

EXAMPLE 5: COMPARISON OF GENES DIFFERENTIALLY EXPRESSED IN BLOOD OF BREAST CANCER PATIENTS AND NORMAL CONTROLS USING HYBRIDIZATION ARRAY ANALYSIS

Messenger RNA expression patterns in cells circulating in the blood of breast cancer patients was compared with normal controls using a membrane-based hybridization array spotted with tags representing 196 different candidate tumor marker genes. The gene tags included 170 DD-identified genes up- and down-regulated in the metastatic breast cancer cell line MDA-MB-435, and 26 literature-reported cancer genes. Blood was drawn from breast cancer patients at the time of chemotherapy treatments or immediately prior to surgery. All donors were female. Age distributions of volunteers and patients were similar. To reduce contamination of samples with skin epithelial cells from the needle stick, 3 mL of blood was drawn into a first tube, which was discarded. 5-10 mL of whole blood was then drawn into a second EDTA-tube. White cells were isolated within 4 hours using a red cell lysis procedure (Ambion Inc., Austin, TX) and cell pellets were stored at –80°C. Total RNA was purified using Trizol reagent (Life Technologies, Rockville, MD). Agarose gel electrophoresis and densitometry determined RNA quality and verified its concentration. Follow-up samples were drawn from three healthy volunteers. Samples N0b and N0c were drawn from the same individual as was initial sample N0a, after intervals of 5 months and 5 months plus one week. N2b and N2c were from the same individual as N2a following intervals of 2 weeks and 5 months. N2c(1) and N2c(2) were run on different days with different membranes using the same N2c RNA preparation. N4b was from the same individual as

N4a following an interval of 2 weeks. N4a(1) and N4a(2) are repeated experiments using the same RNA preparation.

cDNA arrays were prepared (Martin et al 2000 Cancer Research 60:2232) by spotting cDNA onto positively-charged nylon membranes (Micron Separations Inc., Westboro, MA). Fourty replicate membranes were prepared. Radiolabeled cDNA probes were prepared (Fournier et al, 2000, Methods in Molecular Biology, Humana Press, Totowa, NJ, in press; Martin et al 2000, Cancer Research 60:2232). Membranes were pre-hybridized 3 hours in formamide-based buffer (Fournier et al, 2000, Methods in Molecular Biology, Humana Press, Totowa, NJ, in press) at 41°C. Probe was then added to buffer and membranes were hybridized 18 hours at 41°C. Membranes were washed (Fournier et al, 2000, Methods in Molecular Biology, Humana Press, Totowa, NJ, in press), exposed to phosphor-imaging screens for two days and analyzed using the Storm system and ImageQuant software (Molecular Dynamics). Membranes were stripped and reused three times. Profiles represent individual experiments. Control experiments were performed to test array reproducibility. Repeated analyses of a single preparation of MDA-MB-435 RNA on different days using different membranes showed that 95% of data points fell within 2.5-fold limits. Similar results were obtained with other RNA preparations. Signal intensities were quantified and normalized (Martin et al 2000, Cancer Research 60:2232).

48 blood samples (15 normal volunteers, 26 breast cancer patients) were analyzed on arrays. Cluster analysis allowed identification of a single group of 12 breast cancer marker genes that were expressed at higher levels in the blood of breast cancer patients than that of healthy volunteers. The 12 cancer marker genes as a group were elevated in 77% (10 of 13) of untreated invasive cancer patients. These marker genes were generally not elevated in patients treated with chemotherapy (29%, 2 of 7). The genes were also not generally elevated in patients with local disease (DCIS) 17% (1 of 6). Three false positives initially resulted among 15 healthy volunteers (19%). Follow-up tests of three (N0, N2, and N4) confirmed two originally negative results (N2 and N4) and corrected one originally positive result (N0), bringing false positive rate to 13% (2 of 15). Gene expression levels of a set of 5 "housekeeping" control genes often used as non-differential controls. These included Rib protein S10, GAPGH, Rib protein (36B4), Clatherin It chain A and Cytochrome C oxidase. These genes were generally expressed at non-differential levels in the blood samples.

cDNA arrays also included genes reported in the literature to be useful expression markers for breast cancer, including keratins 14 and 19, muc-1 and her2/neu. Keratins are

specific for epithelial cells. Keratin 14 and 19 distinguish the two epithelial cell types present in breast tissue, myoepithelial and luminal epithelial cells, respectively (Taylor-Papadimitriou et al, 1989, J. Cell Science 94 :403). Keratin 19 has been reported to be a useful marker for disseminated tumor cells in the blood in numerous studies (e.g. Slade et al, 1999, J Clinical Oncology 17 :870), though others have reported elevated keratin 19 in similar percentages of healthy individuals and cancer patients (Grunewald et al, 2000, Lab Investigation 80 :1071). Both keratins were found at elevated levels in the blood of most healthy individuals and cancer patients. Muc-1 and her2/neu have also been reported as useful markers for breast cancer in some blood studies (de Cremoux et al, 2000, Clinical Cancer Research 6 :3117 ; Wasserman et al, 1999, Molecular Diagn. 4 :21) but not others (Berois et al, 2000, European J. Cancer 36:717; Eltahir et al, 1998, British J Cancer 77 :1203). Muc-1 expression was high in normal blood. In contrast, her2/neu was consistently low in healthy individuals, but was elevated in 4 of 13 (31%) untreated invasive breast cancer patients.

Quantitative real-time PCR was used to confirm cDNA array results. Real-time PCR of three markers (mdm-2, gro-alpha, and maspin) was performed in a subset of both cancerous and healthy blood samples using the standard curve method. 2 μg total RNA from white cells were denatured at 70°C for 10 minutes and then reverse-transcribed in 30 μl reaction mixture containing 250 μM of each dNTP, 50 U of reverse transcriptase enzyme (Superscript II, Life Technologies), 80 ng/μl oligo(dT) 12-18 primers (Life Technologies), 1X PCR buffer, 2.0 mM $MgCl_2$ (Applied Biosystems, Foster City, CA) at 42°C for 50 minutes. cDNAs were purified on Sepharose G-50 columns (Boehringer Mannheim, Indianapolis, IN), dried and resuspended in 50 μl of dH_2O. Reactions omitting enzyme or RNA were used as negative controls. Specific primers for human maspin, mdm-2 and gro-alpha genes were designed to work in the same cycling conditions (95°C for 10 minutes followed by 45 cycles at 95°C for 15 seconds and 60°C for 1 minute) generating products with sizes 100 to 150 bp. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), β-actin, ribosomal protein PO and cyclophilin were tested as reference genes using a group of four normal blood samples and four patients' blood. GAPDH and ribosomal protein PO levels were similar in tested samples (and consistent with cDNA amounts as determined by OD_{260} and fluorescence of the single-strand DNA-specific dye OliGreen, Molecular Probes), while beta-actin and cyclophilin showed high variations. PCR was performed using an ABI PRISM 7700 Sequence Detector (Applied Biosystems, Foster City, CA) and SYBR® Green reagents. For each reaction, standard curves for both target and reference gene were made using six 4-fold serial dilutions of N27 cDNA. All samples were run in triplicate.

Relative amounts of maspin, mdm-2, and gro-alpha transcripts were calculated by comparison with standard curves. Data were normalized to GAPDH or the amount of cDNA in the reaction. All samples were resolved in a 1.8% DNA agarose gel to confirm the PCR specificity.

For all three genes tested, real-time PCR confirmed array results that the markers were expressed at higher levels in the blood of cancer patients than that of healthy individuals.

Quantitative real-time PCR was also used to determine if the cDNA arrays were sensitive enough to detect disseminated tumor cells in blood. Maspin expression in the blood of healthy volunteers was at the limit of array detection. Low, but measurable maspin levels were detected by arrays in 33% of healthy volunteers (7 of 21), while the others had undetectable levels.

The absolute level of maspin transcripts was determined by comparison with a standard curve generated from dilutions of quantitated maspin plasmid. Identical standard curves were generated if maspin plasmid was measured alone or added to normal blood cDNA. Measurements of maspin transcripts in blood were normalized to the amount of cDNA in reverse transcription reactions, which was quantitated by two methods: OD_{260} and fluorescence of the single-strand DNA dye OliGreen. Results were adjusted to account for the use of double-stranded plasmid in standard curves and for the size of maspin cDNA (3 Kb), which is longer than an average cellular transcript (2 Kb). Maspin was quantitated in cDNA preparations and its level relative to total cDNA was assumed to be equivalent in cDNA and RNA.

In blood sample N21, a 2 μL reverse transcription reaction contained $2.4x10^{-4}$ pg maspin cDNA and $3.6x10^{4}$ pg total cDNA. Thus, maspin transcripts represented 1 in $2.3x10^{8}$ transcripts (e.g. $2.2x10^{-9}$ maspin messages/total cellular messages). Since a typical mammalian cell has $3.6x10^{5}$ RNAs in its cytoplasm, this corresponds to an *in vivo* level of $1.6x10^{-3}$ copies per white blood cell. Maspin expression was similar in N27, which was not tested by arrays, and 2-fold lower in N12, which had an undetectable level of maspin by arrays. The limit of detection for the arrays was approximately 1 in $2x10^{8}$ transcripts. This is two orders of magnitude less sensitive than PCR, which has a limit of approximately 1 in 10^{10} messages. However, it is three orders of magnitude better than oligonucleotide microarrays with a limit of 1 in $3x10^{5}$ messages (22). The enhanced sensitivity of our arrays can be accounted for by the use of ^{32}P rather than fluorescence labeling, membranes rather than glass, and long cDNA tags rather than oligonucleotides.

The detection limit of 1 in $2x10^{8}$ can be used to calculate whether arrays are capable of detecting low numbers of disseminated tumor cells in blood. If one assumes tumor cells and white cells express similar total amounts of RNA and that the marker used is abundant in

tumor cells (e.g. 1000 copies/cell), then it can be calculated that the arrays can detect as few as 1 in 10^6 cells, e.g. 5 tumor cells per mL of blood. This level of sensitivity is sufficient for the detection of tumor cells in blood. (5,6)

EXAMPLE 6: IDENTIFICATION OF GENE CLUSTERS WITH CLINICALLY RELEVANT EXPRESSION PATTERNS FROM BLOOD OF BREAST CANCER PATIENTS.

Messenger RNA expression levels of 196 candidate tumor marker genes were assayed using hybridization assays in blood of breast cancer patients and normal controls. The statistical significance of individual gene expression levels was tested by permutation analysis and cluster analysis was used to organize the genes so that those with the most closely related expression patterns were positioned adjacent to each other (Eisen et al, 1998, Proc. Natl. Acad. Sci. U.S.A. 95:14863).

The statistical significance of individual gene expression levels was tested by permutation analysis (Cox and Hinkley, Theoretical Statistics, Chapman and Hall, London, UK). The software used, PERMAX, performs permutation 2-sample t tests on large arrays. The significance level for each gene is determined by comparing its statistic to the permutation distribution of the max statistic over all genes. The main use is to determine genes that are most different between two groups. We applied this test to a group of blood samples that included the first sample collected from each of the 15 healthy volunteers and all of 26 breast cancer patients. PERMAX identified 5 "most significant" genes that drive distinction among the 12 genes of cluster I. These genes include CD44, maspin, gro-alpha, tubulin and N33. These 5 genes were found to be most distinguished between cancer patient and healthy volunteer groups. A cross-validation procedure done by dropping one tissue at each test confirmed these 5 most significant genes. Other significant genes ($p<0.05$) identified by PERMAX but not represented in cluster I included β-actin ($p=0.0014$), doc-1 ($p=0.0124$), mac25 ($p=0.0126$), unknown 28/13 ($p=0.0180$), unknown TG90D ($p=0.0188$), desmoglein 2 ($p=0.0240$), c-fos ($p=0.0262$), interferon γ ($p=0.0278$) and chondroitin sulfate proteoglycan ($p=0.0372$).

Cluster analysis is a computer method that calculates correlation coefficients between all gene and tissue data sets. It then organizes the positions of the rows and columns of the display and generates hierarchical trees that indicate the degree of relatedness by the height of the nodes. Data sets were logarithmically transformed. Average linkage hierarchical clustering was performed using an uncentered correlation for both array and gene clustering dimensions. Hierarchical cluster analysis performed using alternate similarity metrics gave similar

conclusions, as did K means clustering.. Hierarchical cluster analysis was used to classify blood samples based on array results. This analysis was performed using expression information from only the 12 Cluster I genes. Cluster analysis sorted blood samples into three classes, A, B and C. Class A included samples with a high percentage of over-expressed (red) Cluster I genes, Class B included samples with a mix of over- and under-expressed (red and blue) genes and Class C included samples with under-expressed (blue) genes. Classification of blood samples was in close agreement with the disease status of the blood donors ($p=0.0022$; Fisher's exact test). Class A was predominantly composed of samples from patients with untreated invasive ductal breast cancer. It included 69% (9 of 13) patients with untreated invasive ductal carcinoma and only 14% (3 of 22) healthy volunteers. Classes B and C together included predominantly healthy volunteers: 86% (19 of 22). Patients with localized disease (DCIS) generally fell into Class B (75%, 3 of 4), indicating that the markers did not effectively detect non-invasive cancer. Patients who had been treated with chemotherapy prior to blood sampling were split between Class A (33%, 2 of 6) and Class C (67%, 4 of 6). This division between cancer-predominant and normal-predominant classes may reflect the extent of treatment efficacy.

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

What is claimed is:

1. A method of diagnosing a neoplasm in a subject, the method comprising:

a) measuring a subject profile of tumor-associated genes; and

b) comparing said subject profile with a reference profile of said tumor-associated genes,

wherein a difference between the subject profile of expression of a tumor-associated genes and the subject profile of expression of said tumor associated gene indicates that said subject suffers from or is at risk of developing a neoplasm.

2. The method of claim 1, wherein said tumor-associated genes are selected from the group consisting of DFCIs: 1-51 and 52.

3. A method of diagnosing a neoplasm in a subject, the method comprising:

a) measuring expression of two or more nucleic acid sequences selected from the group consisting of DFCIs: 26, 28, 33, 36-43 and 44 in a subject derived cell population; to yield a subject profile; and

b) comparing the expression of said nucleic acid sequences to the expression of nucleic acid sequences in a cancer reference profile, wherein a substantial similarity between the expression of nucleic acid sequences in said subject-derived cell population and said cancer reference profile indicates the presence of a neoplasm in said subject.

4. A method of diagnosing a neoplasm in a subject, the method comprising:

a) measuring expression of two or more nucleic acid sequences selected from the group consisting of DFCIs: 11, 12, 45-51 and 52 in a subject derived cell population to yield a subject profile; and

b) comparing the expression of said nucleic acid sequences to the expression of nucleic acid sequences in a cancer reference profile, wherein a substantial similarity between the expression of nucleic acid sequences in said subject-derived cell population and said cancer reference profile indicates the presence of a neoplasm in said subject.

5. A method of diagnosing a neoplasm in a subject, the method comprising:

a) measuring expression of one or more nucleic acid sequences selected from the group consisting of DFCIs: 7, 8, 25 and 31 in a subject derived cell population to yield a subject profile; and

b) comparing the expression of said nucleic acid sequences to the expression of nucleic acid sequences in a cancer reference profile, wherein a substantial similarity between the expression of nucleic acid sequences in said subject-derived cell population and said cancer reference profile indicates the presence of a neoplasm in said subject.

6. A method of assessing the prognosis of a subject with a neoplasm, the method comprising:

a) measuring over time the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population to yield a subject profile to yield a subject profile; and

b) comparing said subject profile to a cancer reference profile, wherein an increase in similarity between said subject profile and said cancer profile over time indicates an adverse prognosis of said subject.

7. A method of assessing the prognosis of a subject with a neoplasm, the method comprising:

a) measuring over time the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population to yield a subject profile to yield a subject profile; and

b) comparing said subject profile to a cancer reference profile, wherein an decrease in similarity between said subject profile and said cancer profile over time indicates an favorable prognosis of said subject.

8. A method of assessing the prognosis of a subject with a neoplasm, the method comprising:

a) measuring over time the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population to yield a subject profile to yield a subject profile; and

b) comparing said subject profile to a non- cancer reference profile, wherein an increase in similarity between said subject profile and said cancer profile over time indicates a favorable prognosis of said subject.

9. A method of assessing the prognosis of a subject with a neoplasm, the method comprising:

a) measuring over time the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population to yield a subject profile to yield a subject profile; and

b) comparing said subject profile to a non- cancer reference profile, wherein an decrease in similarity between said subject profile and said cancer profile over time indicates a adverse prognosis of said subject.

10. A method of assessing the prognosis of a subject with a neoplasm, the method comprising:

a) measuring expression of two or more nucleic acid sequences selected from the group consisting of DFCIs: 26, 28, 33, 36-43 and 44 in a subject derived cell population to yield a subject profile; and

b) comparing said subject profile to a cancer reference profile, wherein an increase in similarity between said subject profile and said cancer profile over time indicates an adverse prognosis of said subject.

11. A method of assessing estrogen receptor status in a subject, the method comprising:

a) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-27 and 28 in a subject derived cell population; and

b) comparing the expression of said nucleic acid sequences to the expression of said nucleic acid sequences in a reference profile comprising at least one cell whose estrogen receptor status is known,

thereby indicating estrogen receptor status in said subject.

12. A method of assessing breast tumor stage in a subject, the method comprising:

a) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-30 and 31 in a subject derived cell population; and

b) comparing the expression of said nucleic acid sequences to the expression of said nucleic acid sequences in a reference profile comprising at least one cell whose breast tumor stage is known,

thereby assessing breast tumor stage in said subject.

13. A method of assessing breast tumor size in a subject, the method comprising:

a) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 29-34 and 35 in a subject derived cell population; and

b) comparing the expression of said nucleic acid sequences to the expression of said nucleic acid sequences in a reference profile comprising at least one cell whose breast tumor size is known,

thereby assessing breast tumor size in said subject.

14. A method of assessing the efficacy of a treatment of a neoplasm in a subject, the method comprising:

a) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population to yield a subject profile; and

b) comparing said subject profile to a cancer reference profile, wherein an increase in similarity between said subject profile and said cancer profile over time indicates the treatment is not efficacious.

15. A method of assessing the efficacy of a treatment of a neoplasm in a subject, the method comprising:

a) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population to yield a subject profile; and

b) comparing said subject profile to a cancer reference profile, wherein an decrease in similarity between said subject profile and said cancer profile over time indicates the treatment is efficacious.

16. A method for identifying a therapeutic agent suitable for treating a neoplasm in a selected subject, the method comprising:

a) contacting a subject derived cell population with a test agent;

b) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in said subject derived cell population; and

c) comparing the expression of said nucleic acid sequences to the expression of said nucleic acid sequences a reference profile,

thereby identifying a therapeutic agent appropriate for said subject.

17. A method of identifying a candidate therapeutic agent suitable for treating a neoplasm, the method comprising:

a) contacting a test cell population with a test agent;

b) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in said test cell population; and

c) comparing the expression of said nucleic acid sequences to the expression of said nucleic acid sequences in a reference profile

thereby identifying a therapeutic agent for treating a neoplasm.

18. A method of categorizing a neoplasm in a subject, the method comprising:

a) measuring the expression two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-51 and 52 in a subject derived cell population; and

c) comparing the expression of said nucleic acid sequences to the expression of said nucleic acid sequences in a reference profile

thereby categorizing said neoplasm in said subject.

19. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17 or 18, wherein said subject derived cell population comprises a breast cell.

20. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17 or 18j, wherein said subject derived cell population comprises a cell found in blood.

21. The method of claim 17, wherein said test cell population comprises a breast cell.

22. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17 or 18, wherein said method comprises comparing expression of two or more of said nucleic acid sequences.

23. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17 or 18, wherein said method comprises comparing expression of five or more of said nucleic acid sequences.

24. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17 or 18, wherein said method comprises comparing expression of ten or more of said nucleic acid sequences.

25. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 14, 15, 16, 17 or 18, wherein said neoplasm is a breast carcinoma.

26. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 12, 13, 14, 15, 16, or 18, wherein said subject is a human subject.

27. The method of claim 16, or 17,wherein said test agent is an anti-estrogen agent.

28. The method of claim 23 wherein said anti-estrogen agent is tamoxifen.

29. The method of claim 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17 or 18, wherein said reference profile comprises a database.

30. A kit comprising a DFCIX detection reagent which binds two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-52 or a gene product encoded by said nucleic acid sequences.

32. An array comprising a nucleic acid which binds two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-52.

33. A plurality of nucleic acid sequences comprising two or more nucleic acid sequences selected from the group consisting of DFCIs: 1-52.

34. An isolated nucleic acid molecule comprising the nucleotide sequence of DFCIs : 7 or a fragment thereof.

35. An isolated nucleic acid molecule comprising the nucleotide sequence of DFCIs : 8 or a fragment thereof.

36. An isolated nucleic acid molecule comprising the nucleotide sequence of DFCIs : 25 or a fragment thereof.

37. An isolated nucleic acid molecule comprising the nucleotide sequence of DFCIs : 31 or a fragment thereof.